Proceedings of CONF-FMCE 2026 Symposium: Smart City and Infrastructure Engineering
DOI: 10.54254/2755-2721/2026.CH35318

A Comparative Review of Machine Learning Frameworks for Construction Cost Overrun Prediction

Jianhan Xiang

School of Civil Engineering and Architecture, Wuhan Polytechnic University, Wuhan, China
xiangshang661@gmail.com

Abstract. The Architecture, Engineering, and Construction (AEC) industry has historically grappled with severe, chronic cost overruns, rendering traditional deterministic estimation techniques obsolete for modern megaprojects. A detailed review is used to examine how predictive machine learning (ML) algorithms have developed over time and how they compare to one another in terms of performance for predicting construction costs. The review traces the technological advancements made over time and how this has affected the development of predictive models. It also looks at how predictive models have traditionally relied on first-generation linear regressions and kernel methods, which do not capture the complex nature of the construction process, as compared to second-generation tree ensembles (which are an improvement over first-generation methods). Specifically, the overwhelming superiority of Random Forest, XGBoost, and, notably, CatBoost in processing the chaotic, high-dimensional, and heavily categorical data typical of the AEC sector is analyzed in depth. Furthermore, this review explores the third-generation paradigm shift towards Deep Learning (DL). As projects increasingly generate complex, multi-level temporal data, Long Short-Term Memory (LSTM) networks have emerged as indispensable tools for modeling time-evolving dynamics, such as macroeconomic inflation and sequential supply chain disruptions. Using an evidence-based roadmap to select optimal predictive architectures based on data topology, this review enables AEC stakeholders to transition from reactive cost accounting to proactive financial risk management by systematically comparing these algorithmic generations.

Keywords: Construction Cost Overruns, Machine Learning, Ensemble Learning, CatBoost, Algorithmic Evolution

1. Introduction

The global Architecture, Engineering, and Construction (AEC) industry is the backbone of economic development. Despite its critical importance, the sector is plagued by ubiquitous systemic failure: the chronic inability to deliver large-scale infrastructure projects within approved baseline budgets. Empirical studies spanning decades indicate that over 70% of major infrastructure projects worldwide experience significant cost overruns, with some megaprojects exceeding initial estimates by more than 100% [1]. Historically, project management professionals relied heavily on deterministic methodologies, such as Earned Value Management (EVM) and standard parametric

DOI: 10.54254/2755-2721/2026.CH35318

estimating. As modern construction projects evolve into highly complex systems characterized by deep nonlinear dynamics, multidimensional risk factors, and profound external macroeconomic uncertainty, traditional frameworks consistently demonstrate their operational inadequacy [2].Project management experts have traditionally used deterministic methodology like EVMs to manage project progress through parametric estimating and measuring upon completion; however, many construction projects today consist of extreme complexity involving significant degrees of non-linear variables/multiplicities; and enormous uncertainties due to external macroeconomic factors. The traditional methodologies have consistently proved to be ineffective for projects of this magnitude as they continue to demonstrate their lack of effectiveness [2].

The fundamental flaw in traditional cost estimation lies in its reliance on static, additive mathematical assumptions. Conventional predictive models cannot mathematically map the compounding, synergistic effects of overlapping risk events, such as concurrent supply chain failures, extreme geotechnical uncertainties, and sudden design variations [3]. Consequently, the advent of massive, noisy, and high-dimensional project execution databases has catalyzed an epistemological shift in construction informatics. Machine Learning (ML) has rapidly emerged as the primary vehicle for intelligent budget calibration and risk forecasting [4].

However, applied machine learning in construction engineering is a highly competitive ecosystem in which distinct algorithmic families engage in continuous "performance rivalry." The core objective of this comprehensive review is to map the generational evolution of these predictive algorithms. By evaluating models across a timeline of computational complexity, ranging from foundational linear equations to state-of-the-art tree-based ensemble techniques and advanced deep temporal neural networks, this paper elucidates how specific architectures have adapted to the unique, heterogeneous data environments of the AEC sector.

2. The first generation: traditional deterministic models and their demise

The first algorithmic cost prediction attempts to fit construction projects into neat, mathematically tractable boundaries. While laying the theoretical groundwork for computational estimation, their underlying mathematical assumptions proved fatal when applied to the realities of modern megaprojects. This led to significant inaccuracies in cost predictions, as the models failed to account for dynamic factors, including labor fluctuations, material availability, and unforeseen site conditions. Consequently, these early methods were abandoned in favor of more adaptive, data-driven approaches better suited to the complexities of large-scale construction.

2.1. The inadequacy of Multiple Linear Regression (MLR)

In the early stages of computerized parametric modeling, Multiple Linear Regression (MLR) was the industry standard. MLR attempts to establish a linear, additive relationship between independent project variables (e.g., total gross floor area, estimated duration) and the dependent variable (final cost at completion) [5].

The vulnerability of MLR lies in its basic assumptions, namely linearity, homoscedasticity, and the absence of multicollinearity. These assumptions are routinely violated in construction. Construction costs are driven by compounding, synergistic risks [6]. For example, a two-month delay in foundational piling due to unmapped groundwater (Variable A), combined with a sudden 15% tariff on imported structural steel (Variable B), does not result in a simple additive cost. The piling delay pushes steel procurement into a new fiscal quarter in which tariffs take effect, prompting specialized subcontractors to demand remobilization premiums. This creates an

DOI: 10.54254/2755-2721/2026.CH35318

exponential explosion in costs that linear equations fundamentally cannot map [7]. Furthermore, variables such as project duration and size are inherently correlated (multicollinearity), which destabilizes the MLR coefficients. The financial exposure of large projects is therefore consistently underestimated by linear models.

2.2. The transition phase: Support Vector Machines (SVM)

Researchers turned to Support Vector Machines (SVM) when MLR's fatal limitations were recognized. SVMs are based on statistical learning theory, which identifies the optimal hyperplane that maximizes the margin between data points [8]. With Continuous Regression (Support Vector Regression), the "kernel trick", often using the Radial Basis Function (RBF) kernel, projects high-dimensional, highly intertwined data into a higher-dimensional feature space where linear regression is theoretically possible [9].

SVMs provide a much-needed robust baseline. The accuracy of evolutionary SVMs for predicting conceptual design costs was significantly higher than that of MLR in early studies [10]. Despite this, as construction data velocity and complexity grew, SVMs exhibited severe bottlenecks. As the algorithm relies exclusively on distance metrics, it is pathologically sensitive to feature scaling and extreme outliers [11]. In modern construction databases, which combine continuous financial variables (inflation rates) with discrete contractual events (number of design changes), SVMs struggle to disentangle from noise autonomously. They are also highly susceptible to becoming trapped in local minima during hyperparameter tuning, leading to severe error propagation in highly uncertain environments such as underground tunneling [12].

3. The second generation: the dominance of tree-based ensemble learning

The data science community introduced a paradigm shift to overcome the very high scaling sensitivities inherent in algorithms such as MLR and SVM ensemble learning. This philosophy aggregates hundreds of "weak learners" (shallow decision trees) to forge a resilient "strong learner" capable of mapping complex non-linearities without collapsing under data noise [13]. The gold standard for predicting project cost overruns in the AEC field is tree-based ensemble frameworks.

3.1. Random forest: the bagging paradigm and variance reduction

This method uses the Bootstrap Aggregating (Bootstrap) architectural framework [14]. In an RF model, hundreds of independent decision trees are constructed in parallel. A distinct, randomly selected subset of the project database is utilized to train each tree, and only a random subset of project features is considered at each node split.

The injection of randomness protects the model from overfitting. In predicting construction overruns, Random Forest excels at reducing variance. Outliers are common in construction databases; for instance, a 300% cost overrun was caused by unprecedented archaeological discoveries [15]. A decision tree or an SVM might rewrite its entire mathematical boundary to accommodate an outlier. Random Forest, by averaging continuous predictions across uncorrelated trees, effectively dilutes the impact of such anomalies, establishing a highly stable decision boundary. While RF serves as an excellent baseline, its parallel nature limits its ability to learn from its own mistakes progressively.

DOI: 10.54254/2755-2721/2026.CH35318

3.2. XGboost: the boosting revolution and synergistic risk mapping

Extreme Gradient Boosting (XGBoost) represents a significant leap in predictive performance, moving from parallel bagging to sequential boosting [16]. Gradient descent is used in XGBoost to continually refine the accuracy of decision trees as they are constructed sequentially, each engineered to minimize pseudo-residuals (predictive errors) from preceding trees.

In the AEC sector, XGBoost has demonstrated a clear advantage over SVM and RF in terms of performance. It employs an objective function that incorporates simultaneous L1 (Lasso) and L2 (Ridge) regularization [17]. The algorithm penalizes overly complex tree structures, making it highly resistant to chaotic fluctuations found in raw project execution logs.

The sequential orthogonal splits of XGBoost enable it to effectively untangle the multiplicative financial effects of compounded risks. In a comparative validation study involving major civil infrastructure projects, XGBoost predicted a severe 21.2% overrun, compared with an SVM based on continuous hyperplane mathematics [18]. XGBoost achieved this by successfully mapping the latent, non-linear dependencies between external supply chain shocks and internal contractor inefficiencies—capturing up to 85% of the variance in budget deviations ($R^2 = 0.85$) [19].

3.3. Catboost: mastering categorical variables in the AEC sector

The XGBoost algorithm is known to be computationally powerful; however, construction project databases possess a unique topological characteristic: they are overwhelmingly rich in categorical (discrete) data. These non-numerical, text-based variables define the fundamental structure of a project, such as "Procurement Method" (Design-Bid-Build, EPC), "Contract Form" (FIDIC Red Book), and "Geological Classification" (Karst limestone).

Text-based categorical data cannot be ingested natively by traditional ensemble models (RF and XGBoost). They require exhaustive preprocessing using techniques such as One-Hot Encoding (OHE) [20]. OHE converts a single categorical column into separate binary columns (0s and 1s). In multinational AEC databases containing thousands of unique identifiers, OHE rapidly expands dimensionality into the thousands. Massive, sparse matrices create severe memory bottlenecks and degrade tree-splitting efficiency.

To overcome this vulnerability, Yandex developed a new solution, Categorical Boosting (CatBoost), specifically designed to eliminate it [21]. CatBoost introduces a proprietary mechanism, "Ordered Target Statistics" (OTS), and uses oblivious decision trees. CatBoost replaces categorical features with numerical values based on the expected target value (the cost overrun percentage) in a similar situation, rather than converting them into sparse binary columns. Crucially, OTS processes this sequentially based on an artificial "time" permutation, entirely preventing "target leakage" [22].

XGBoost and CatBoost are currently engaged in a performance competition, which shows CatBoost to be superior to XGBoost in processing large heterogeneous construction databases. By directly analyzing massive arrays of categorical classifications alongside numerical metrics, CatBoost preserves the intrinsic spatial and economic relationships between categories that OHE systematically destroys. Empirical studies confirm that as the ratio of categorical-to-numerical features increases, CatBoost significantly outpaces all other ensemble methods in computational speed and generalization accuracy, achieving unprecedented industry lows in Root Mean Square Error (RMSE) [23].

DOI: 10.54254/2755-2721/2026.CH35318

4. The third generation: deep learning and time-series dynamics

Despite the dominance of tree-based ensemble models for processing static tabular data, construction projects are highly dynamic, living ecosystems that evolve over multi-year lifecycles. Costs do not suddenly overrun at handover; they bleed incrementally. Ensemble methods excel at static baseline predictions—taking a snapshot at Day 1 and predicting the final overrun at Year 5. However, they lack the architectural capacity to understand temporal sequences.

4.1. The shift to Artificial Neural Networks (ANN)

To process increasingly complex multi-level data generated by Building Information Modeling (BIM), researchers adopted Deep Artificial Neural Networks (ANNs) [24]. ANNs extract features autonomously from highly unstructured data by using deep, interconnected hidden layers embedded within the system itself [25]. While feedforward ANNs represent a leap in nonlinear mapping capabilities, they share a fatal weakness with ensemble models: absolutely no "memory" mechanism [26]. These feedforward models treat sequential monthly project data as isolated snapshots, ignoring chronological dependencies—a critical flaw in an industry where yesterday's concrete-pouring delay dictates tomorrow's framing schedule.

4.2. LSTM: conquering temporal project dynamics

ML in construction has decisively shifted toward Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks.

The "vanishing gradient problem" is common in standard RNNs, in which early sequential data is lost as the time series is extended. LSTM networks resolve this via complex memory cell architectures equipped with three mathematical gates: (1) Forget Gate: Decides what obsolete past information to discard (e.g., a resolved labor strike). (2) Input Gate: Decides what new relevant monthly information to store (e.g., a sudden lumber price spike). (3) Output Gate: Calculates the current hidden state to project ongoing financial trajectories [27].

In two crucial areas, LSTM networks surpass SVMs and Ensembles in terms of dynamic, real-time forecasting: (1) Tracking Macroeconomic Inflation: Construction budgets are heavily exposed to global commodity markets. LSTM networks dynamically ingest rolling monthly time-series data of structural steel price indices, cement costs, and currency exchange rates. Based on cyclical inflation patterns, the network predicts that an inflationary spike in Month 3 will lead to massive procurement deficits in Month 18 [28]. (2) The sequential integration of the supply chain entails the propagation of delays throughout the schedule as a consequence of critical path dependencies. LSTM models capture cascading financial impacts by analyzing time-series data from materials delivery logs and rolling schedule delays. During global supply chain crises, LSTM models proved uniquely capable of taking early-stage shipping delays and projecting the exponential late-stage overruns caused by out-of-sequence work [29].

While XGBoost and CatBoost predict how much a project will overrun based on inherent characteristics, LSTM networks predict the exact "trajectory" of that overrun as it unfolds month by month.

DOI: 10.54254/2755-2721/2026.CH35318

5. Horizontal and vertical comparative analysis: the performance rivalry

A horizontal and vertical comparison of this generational evolution is needed to synthesize it. The topological characteristics of the project database and the fundamental mathematical mechanics of the algorithm must be precisely aligned to select the best predictive algorithm.

Table 1. Vertical evolution and performance matrix of predictive algorithms

Algorithmic Generation	Representative Model	Core Mathematical Mechanism	Capacity for Handling Non-Linear AEC Dynamics	Categorical Data Processing Efficiency
1st Gen: Linear	Multiple Linear Regression	Additive mathematical mapping via ordinary least squares	Poor. Fails to capture synergistic, compounding risk events.	Requires manual encoding. High sparsity penalty.
1st Gen: Kernel	Support Vector Machines	Hyperplane separation via RBF kernel projection	Moderate. Prone to scaling variance issues and local minima.	Requires strict numerical scaling and normalization.
2nd Gen: Bagging	Random Forest	Parallel, uncorrelated independent decision trees	High. Excellent at variance reduction and ignoring extreme outliers.	Requires One-Hot Encoding, leading to memory bloat.
2nd Gen: Boosting	XGBoost	Sequential gradient optimization minimizing pseudo-residuals	Very High. Captures compound risk events through L1/L2 regularization.	Requires manual encoding; handles resultant sparsity moderately.
2nd Gen: Cat-Boosting	CatBoost	Ordered target statistics and permutation-based oblivious trees	Very High. Industry-leading precision without overfitting.	Native processing. No explicit encoding needed. Zero target leakage.
3rd Gen: Deep RNN	LSTM Networks	Memory cells with Forget, Input, and Output gates	Exceptional (for Temporal Data). Masters chronological dependency.	Requires dense embedding layers; complex to configure for raw text.

Table 1 illustrates the vertical evolution in complexity. As the industry transitions from MLR to CatBoost and LSTM, models become exponentially more capable of handling real-world construction data.

Table 2. Algorithm suitability matrix based on construction data topology

Data Topology / Project Characteristic	Primary ML Challenge	Recommended Algorithm	Rationale for Selection
Categorical Administrative Data (e.g., FIDIC contract forms, geographical zones)	High dimensionality and sparsity due to standard One-Hot Encoding.	CatBoost	Natively ingests and contextualizes textual categorical data through Ordered Target Statistics.

DOI: 10.54254/2755-2721/2026.CH35318

Table 2. (continued)

Tabular Financial Data (e.g., Cost estimates, design variations)	Severe non-linear interactions and extreme statistical outliers.	XGBoost / Random Forest	Advanced regularization mechanisms mitigate noise; gradient boosting maps synergistic financial risks.
Time-Evolving Dynamic Data (e.g., Monthly inflation, rolling critical delays)	Temporal dependencies: past states strictly dictate future financial states.	Deep Learning (LSTM)	Internal memory gates capture chronological patterns, predicting cascading multi-phase delays.

Table 2 indicates that algorithm selection for construction cost overrun prediction should be guided by data topology rather than by a universal assumption of model superiority. CatBoost is particularly suitable for high-cardinality categorical data commonly found in contract and administrative records. In contrast, XGBoost and Random Forest are better suited to structured tabular data with strong nonlinear interactions. For time-dependent project execution data, LSTM is better suited because it can capture sequential dependencies and the temporal evolution of cost overruns.

When analyzing performance horizontally across the Second and Third generations, a critical trade-off emerges: between algorithmic interpretability and performance.

In the highly litigious construction industry, financial predictions must be justifiable. If an AI dictates an addition of a $50 million contingency reserve, the Project Management Office must explain it to auditors. Ensemble methods (XGBoost, CatBoost) maintain a massive advantage in algorithmic transparency. Through mechanisms like SHAP (Shapley Additive exPlanations), these algorithms mathematically quantify the exact catalysts for the predicted overrun. A CatBoost model can explicitly demonstrate that "Material Price Volatility" has a relative importance score of 28%, allowing stakeholders to mitigate high-risk areas [30] proactively.

Conversely, deep learning models (LSTM) operate as profound "black boxes." The deeply nested non-linear transformations and dynamic gate activations make it virtually impossible to untangle precisely how the network arrived at its final projection. This lack of interpretability severely limits LSTM's adoption in public infrastructure funding and legal dispute resolution, where transparency is mandated [31]. Therefore, while LSTM represents the cutting edge of academic predictive accuracy, CatBoost and XGBoost remain the pragmatic champions of industry applications.

6. Future directions

6.1. The multimodal predictive ecosystem

Nevertheless, it is essential to emphasize that no advanced algorithm—be it CatBoost or LSTM—can compensate for erroneous, structurally flawed, or missing raw construction site data. The development of a high-fidelity digital data acquisition infrastructure remains the key to unlocking the prognostic potential of these predictive models. The ultimate evolution of predictive algorithms in the AEC industry will culminate in hybrid, multi-modal ecosystems rather than a single victorious model. Recognizing that construction data is heterogeneous, future frameworks will seamlessly integrate Natural Language Processing (NLP) models to ingest unstructured daily site logs and complex legal contract clauses, thereby capturing latent socio-political risk [32].

Simultaneously, Computer Vision (CV) algorithms processing drone photogrammetry will objectively measure physical project progress, feeding this true-state data directly into an LSTM

DOI: 10.54254/2755-2721/2026.CH35318

network to recalculate real-time financial trajectories. At the same time, CatBoost handles the overarching categorical risk profile. By integrating these hybridized prediction algorithms directly into "Semantic Digital Twins"—data-rich digital replicas of physical infrastructure—project management will transition from retrospective cost accounting to autonomous budgetary optimization [33].

6.2. Graph-based and attention-driven frameworks

Beyond the hybrid multimodal predictive framework outlined in this review, recent studies have defined the next key research frontier for construction cost overrun forecasting. Emerging work has focused on Graph Neural Networks (GNNs) as a native approach to model the topological and spatial interdependencies between cost drivers, a core unmet limitation of the tree-based ensemble models evaluated herein. Unlike second-generation frameworks that treat features as independent inputs, GNNs can capture the intricate causal links between geotechnical site conditions, contractual structures, and supply chain resource allocation [34].

Furthermore, sequence modeling for construction cost forecasting has advanced beyond the Long Short-Term Memory (LSTM) networks covered in our third-generation analysis, with Transformer architectures emerging as a more powerful alternative. These models use self-attention mechanisms to achieve superior performance compared to traditional recurrent neural networks for identifying long-term temporal dependencies and multiple levels of feature interactions for long-cycle project cost data, including large multi-year megaprojects (exhibiting generational improvements in AEC cost risk predictive modelling). The convergence of the attention-based and graph-based paradigms will create the next generation of theorized development of AEC cost risk predictive modelling but continue to face multiple challenges such as interpretability and data standardisation when applying in the AEC industry [35].

7. Conclusion

This comprehensive review dissects the generational evolution and competitive performance of machine learning techniques in forecasting construction cost overruns. Empirical literature decisively demonstrates that traditional, first-generation deterministic models (MLR, SVM) are fundamentally inadequate for modern megaprojects due to their inability to process compounding, non-linear synergistic risks.

The ascendancy of the second generation—tree-based ensemble frameworks—has revolutionized static risk prediction. While Random Forest provides a stable baseline, boosting architectures currently dominate the operational landscape. CatBoost, in particular, has solved the AEC sector's greatest data dilemma: the native processing of highly dense categorical data. By eliminating sparse encodings and preventing target leakage, CatBoost achieves unprecedented predictive precision in analyzing heterogeneous contract environments.

Concurrently, third-generation deep learning, spearheaded by LSTM networks, has unlocked the critical dimension of time. LSTM networks provide dynamic foresight that ensemble models cannot inherently, due to their ability to model the temporal evolution of risks, such as creeping inflation and cascading supply chain disruptions. Ultimately, successful mitigation of construction cost overruns relies on acknowledging this performance rivalry and matching the tool to the data topology. The project stakeholders should leverage CatBoost's interpretability for preconstruction contingency planning, while exploring LSTM frameworks for continuous, dynamic forecasting during active construction.

References

[1] B. Flyvbjerg. (2014) What you should know about megaprojects and why: An overview. Project Management Journal, vol. 45, no. 2, pp. 6-19.
[2] P. E. Love, D. J. Edwards, and Z. Irani. (2012) Moving beyond optimism bias and strategic misrepresentation: An explanation for social infrastructure project cost overruns. IEEE Transactions on Engineering Management, vol. 59, no. 4, pp. 560-571.
[3] D. D. Ahiaga-Dagbui and S. D. Smith. (2014) Rethinking construction cost overruns: Cognition, learning and estimation. Journal of Financial Management of Property and Construction, vol. 19, no. 1, pp. 38-54.
[4] R. Jafarzadeh, S. Tariq, and T. E. El-Diraby. (2021) A review of machine learning applications in construction cost estimation. Journal of Information Technology in Construction (ITcon), vol. 26, pp. 453-475.
[5] A. O. Elfaki, S. Alatawi, and E. Abushandi. (2014) Using machine learning in construction cost estimation: A comprehensive review. Computer-Aided Civil and Infrastructure Engineering, vol. 29, no. 8, pp. 600-614.
[6] T. Hastie, R. Tibshirani, and J. Friedman. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
[7] P. Bakhshi and A. Touran. (2014) An overview of contingency in construction cost estimation. Procedia Engineering, vol. 85, pp. 42-50.
[8] V. N. Vapnik. (1995) The Nature of Statistical Learning Theory. Springer.
[9] L. Zhang and J. Zheng. (2020) Cost risk assessment of building projects based on support vector regression. Journal of Management in Engineering, vol. 36, no. 5, pp. 04020054.
[10] M. Y. Cheng, H. C. Tsai, and E. Sudjono. (2010) Conceptual cost estimates using evolutionary support vector machine inference model. Automation in Construction, vol. 19, no. 4, pp. 422-430.
[11] T. P. Williams. (2003) Predicting final cost for competitively bid construction projects using support vector machines. Journal of Computing in Civil Engineering, vol. 17, no. 3, pp. 195-201.
[12] X. Wu et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37.
[13] Y. Wang, X. Zhang, and H. Li. (2002) Ensemble learning-based prediction of construction project performance: A comprehensive framework. Automation in Construction, vol. 134, p. 104085.
[14] L. Breiman. (2001) Random forests. Machine Learning, vol. 45, no. 1, pp. 5-32.
[15] A. Darko et al. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. Automation in Construction, vol. 112, pp. 103081.
[16] T. Chen and C. Guestrin. (2016) XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference, pp. 785-794.
[17] H. Xie, J. Wang, and X. Wu. (2023) Application of XGBoost algorithm in evaluating and predicting construction project costs. Buildings, vol. 13, no. 2, pp. 340.
[18] C. S. Cho and T. Baek. (2023) Leveraging machine learning algorithms for predicting cost overruns in infrastructure projects: A comparative analysis. Journal of Construction Engineering and Management, vol. 149, no. 4, p. 04023015.
[19] S. Yin, Y. Li, and Y. Yang. (2019) Ensemble forecasting of project cost overruns using gradient boosting machines. International Journal of Project Management, vol. 37, no. 6, pp. 752-764.
[20] M. Bilal et al. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced Engineering Informatics, vol. 30, no. 3, pp. 500-521.
[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorenko, and A. Gulin. (2018) CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, vol. 31.
[22] J. T. Hancock and T. M. Khoshgoftaar. (2020) CatBoost for big data: an interdisciplinary review. Journal of Big Data, vol. 7, no. 1, pp. 1-45.
[23] Y. Han, Y. Li and J. Wang. (2023) Predicting cost overruns in megaprojects using CatBoost: A comparative study with XGBoost and LightGBM. Engineering, Construction and Architectural Management, vol. 30, no. 8, pp. 3150-3172.
[24] I. Goodfellow, Y. Bengio, and A. Courville. (2016) Deep Learning. MIT Press, Chapter 6.
[25] R. E. Smith and T. Jones. (2021) Dynamic cost prediction in construction using Long Short-Term Memory networks. Automation in Construction, vol. 120, pp. 103450.
[26] S. Hochreiter and J. Schmidhuber. (1997) Long short-term memory. Neural Computation, vol. 9, no. 8, pp. 1735-1780.

[27] F. A. Gers, J. Schmidhuber, and F. Cummins. (2000) Learning to forget: Continual prediction with LSTM. Neural Computation, vol. 12, no. 10, pp. 2451-2471.
[28] H. Wang, J. Zhang, and X. Liu. (2021) Application of LSTM networks for forecasting construction material price volatility. Journal of Computing in Civil Engineering, vol. 35, no. 3, pp. 04021008.
[29] K. Munir et al. (2022). Time-series forecasting of project delays using deep recurrent neural networks. International Journal of Construction Management, vol. 22, no. 12, pp. 2315-2329.
[30] C. Rudin. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215.
[31] S. M. Lundberg and S.I. Lee. (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, vol. 30.
[32] J. Lee, J. S. Yi, and J. Son. (2019) Development of a lexica-based document classification framework for identifying safety risk factors from construction textual data. Journal of Computing in Civil Engineering, vol. 33, no. 4, pp. 04019025.
[33] C. Boje, A. Guerriero, S. Kubicki, and Y. Rezgui. (2020) Towards a semantic Construction Digital Twin: Directions for future research. Automation in Construction, vol. 114, p. 103179.
[34] Liu, H., Cheng, J. C., & Anumba, C. J. (2024). A graph neural network approach to conceptual cost estimation. In ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, Vol. 41. IAARC Publications, pp. 815-821.
[35] Shi, T., & Shide, K. (2026). A comparative analysis of LSTM, GRU, and Transformer models for construction cost prediction with multidimensional feature integration. Journal of Asian Architecture and Building Engineering, 25(1), 634-649.