Journal of Industrial and Management Optimization

JIMO, 22(10): 5074–5105.
DOI: 10.3934/jimo.2026175
Received: 27 July 2026
Revised: 01 September 2026
Accepted: 10 September 2026
Published: 17 September 2026

https://www.aimspress.com/journal/jimo

Research article

A constraint-coordinated hierarchical catboost framework for structurally consistent multi-target cost prediction in substation engineering projects

Tianqiong Chen[1], Hongda Chen[2], Jiaxiang Wen[3,*], Huijuan Huo[1], Zhikai Yang[3], Jing Duan[1], Cheng Xin[1], Zihao Sun[1] and Jiyuan Zhang[4]

[1] State Grid Economic and Technological Research Institute Co., Ltd., Beijing 102209, China
[2] State Grid Sichuan Power Company, Chengdu 610041, Sichuan Province, China
[3] School of Economics and Management, North China Electric Power University, Beijing 102206, China
[4] State Grid Sichuan Economic Research Institute, Chengdu 610000, China

* **Correspondence:** Email: 120252206078@ncepu.edu.cn.

Abstract: Accurate and structurally consistent cost prediction is essential for investment control, budget review, and refined management in substation engineering projects. Although machine-learning models have shown strong potential for engineering cost estimation, conventional single-stage prediction models usually treat multiple cost indicators as independent regression targets and may ignore the accounting identities among cost components. In this study, we proposed a constraint-coordinated hierarchical CatBoost framework for multi-target cost prediction using 2820 real substation project records collected from the State Grid system in 2024. The architecture decomposed the prediction task into a structured representation of heterogeneous engineering attributes, intermediate-variable completion, target-wise cost prediction, and reliability-weighted engineering-constraint coordination, with the final-account reserve fund treated separately as a sparse residual-type accounting target. The experimental results showed that the proposed framework achieved a global R^2 of 0.9521 and a global wMAPE of 17.06%, outperforming six conventional benchmark models. Although it did not uniformly surpass a strong single-stage CatBoost baseline in all global statistical metrics, it substantially reduced engineering-identity residuals and improved consistency between disaggregated cost components and aggregate investment indicators. These results indicate that the

framework balances statistical accuracy with engineering accounting consistency and can support machine-learning-assisted cost estimation and decision support in power infrastructure projects.

Keywords: substation engineering; engineering cost prediction; CatBoost; hierarchical modeling; engineering constraints; applied machine learning
Mathematics Subject Classification: 68T05, 90C20

1. Introduction

The digital transformation of power infrastructure has increased the demand for reliable, artificial-intelligence-assisted, data-driven decision support in planning, construction, investment control, and asset management. Substation projects are key nodes in transmission and distribution systems, and their investment level directly affects grid expansion, capacity upgrading, and the economic efficiency of power-system development. In practical project management, accurate cost prediction supports budget preparation, design review, investment control, and final-account evaluation. However, substation engineering cost is not determined by a single technical parameter; it is jointly influenced by voltage level, transformer capacity, site conditions, building scale, equipment configuration, civil engineering quantities, installation workload, regional construction conditions, and management-related factors. Therefore, substation cost prediction is a representative applied machine-learning problem in engineering, involving heterogeneous data, nonlinear relationships, multiple coupled targets, and explicit engineering constraints.

Studies have shown that data-driven methods can effectively support complex decision-making tasks in modern power and energy systems. For example, data-driven and physically informed approaches have been applied to power grid dispatch decision-making [1], mixed-generation power-system planning toward high renewable-energy penetration [2], and embodied carbon assessment of power infrastructure construction [3]. At the substation and power-system levels, machine-learning methods have also been used for renewable electricity production prediction [4], flexible load disaggregation [5], electric load forecasting based on deep ensemble learning [6], hydroelectricity generation forecasting using tree-based machine-learning models [7], and photovoltaic efficiency prediction with explainable artificial intelligence techniques [8]. These studies indicate that machine learning can support engineering decisions beyond improving forecast accuracy, particularly when multi-source data, physical logic, and operational constraints must be considered together. This observation is highly relevant to the cost prediction of substation projects, where the model should not only fit historical data but also generate outputs that are interpretable and consistent with engineering accounting logic.

Engineering cost prediction has also attracted increasing attention in infrastructure and construction management. Researchers have investigated power engineering cost estimation using neural networks [9], clustering and denoising strategies for power grid project cost data [10], construction cost prediction based on deep learning [11], hybrid models for tunnel construction duration and cost prediction [12], and transparent construction cost prediction using advanced machine learning and explainable artificial intelligence [13]. In Applied Sciences, related studies have further explored early-stage highway construction cost estimation using machine-learning models [14], stacking heterogeneous ensemble learning for building construction project cost prediction [15], and machine-learning-based cost estimation for AI training-data construction under heterogeneous workflows and cost-driver structures [16]. These works demonstrate that intelligent algorithms can improve the efficiency of cost estimation and provide useful support for early-stage decision-making. Nevertheless, most researchers focus on single-output prediction, general construction projects, or relatively independent cost indicators. In contrast, substation project cost data are usually organized as a system of related cost items, such as construction cost, equipment procurement cost, installation cost, other cost, reserve fund, and static investment. These targets are not independent. Related studies on multi-output regression and forecast reconciliation have likewise emphasized the importance of exploiting inter-target dependencies and maintaining coherence among structurally related outputs [17,18]. Some targets follow additive accounting identities, whereas sparse or reserve-related items may have weak statistical regularity. As a result, directly applying a flat regression model to all cost items may lead to predictions that are numerically accurate in some metrics but structurally inconsistent from an engineering perspective.

From an applied machine-learning perspective, machine-learning models have been widely used in energy forecasting and applied engineering prediction. Hybrid deep-learning models have been developed for electric-vehicle charging demand forecasting [19], photovoltaic power prediction [20–22], and short-term load forecasting [23,24]. Ensemble learning and boosted-tree models have also shown strong performance in renewable-energy forecasting [25], medium- and long-term load forecasting [26], offshore wind power prediction with interpretability analysis [27], and broader machine-learning-based power forecasting systems [28]. Applied Sciences studies also confirm the effectiveness of CatBoost, XGBoost, LightGBM, and stacking-based ensemble learning in applied prediction tasks, including PV efficiency estimation [8], hydroelectricity generation forecasting [7], electric load forecasting [6], and construction cost prediction [15]. Among these methods, CatBoost is particularly suitable for structured tabular data because it can handle heterogeneous features, categorical variables, nonlinear interactions, and complex feature combinations. It has been successfully combined with residual correction [20], multi-lag feature engineering [23], stacking strategies [25], double-layer structures [26], interpretability techniques [27], and metaheuristic hyperparameter optimization in different engineering and energy applications [29–31].

However, directly transferring machine-learning methods and modeling frameworks to substation cost prediction remains challenging. First, the input features of substation projects are highly heterogeneous, covering categorical, numerical, Boolean, geographic, voltage-level, temporal, and engineering-quantity variables. Second, the output side contains multiple cost targets with strong coupling relationships. These targets reflect different aspects of the same project, so they should be predicted in a coordinated manner rather than as isolated regression tasks. Third, some cost indicators exhibit zero inflation, sparsity, or irregular fluctuation. For these targets, conventional regression models may produce unstable percentage errors and weak explanatory performance. Fourth, engineering cost prediction has an inherent structural requirement: The predicted subitems should be consistent with aggregate investment indicators. This requirement is particularly important in budget review and final-account acceptance, where cost items are not merely statistical labels but also accounting components constrained by engineering identities.

To address these issues, we propose an engineering-constraint-aware hierarchical CatBoost framework for multi-target cost prediction in substation engineering projects. We use 2820 real substation project records collected from the State Grid system in 2024. After data cleaning, mixed-type preprocessing, and layered feature organization, the prediction task is decomposed into four connected stages: Basic feature input, intermediate-variable completion, target cost prediction, and engineering-constraint coordination. In the first stage, basic engineering attributes are transformed into a structured feature representation. In the second stage, intermediate engineering variables are completed from the basic inputs, so that the model can make use of implicit engineering quantities that are not always directly available. In the third stage, target-wise CatBoost models are trained to predict major cost indicators. In the fourth stage, structural coordination weighted by out-of-fold residual variance is introduced to reduce violations of predefined engineering accounting identities. In addition, a derivation strategy is used for selected sparse or identity-related targets to avoid unstable direct regression.

The proposed framework differs from conventional single-stage prediction models in modeling logic and application objective. A single-stage model directly maps basic project attributes to target costs, which may be effective for statistical fitting but does not explicitly represent the layered formation process of engineering cost. By contrast, the proposed framework organizes the prediction process according to the relationship among basic attributes, intermediate engineering quantities, final cost components, and aggregate investment indicators. This design is not intended to simply increase model complexity. Instead, we aim to improve the practical applicability of machine-learning prediction by incorporating engineering structure into the modeling workflow. Therefore, the framework emphasizes not only predictive accuracy, but also structural closure, output feasibility, and interpretability for engineering cost management.

The main contributions of this study are as follows:

First, we establish a real-world multi-target cost prediction task for substation engineering projects based on actual budget-review and final-account records. Unlike simulated samples or general public construction datasets, the dataset used in this study reflects realistic cost formation characteristics in power infrastructure projects, including heterogeneous engineering attributes, coupled cost components, sparse reserve-related indicators, and accounting relationships among disaggregated and aggregate cost items. This provides a practical data foundation for evaluating applied machine-learning models in substation cost management.

Second, we propose a constraint-coordinated hierarchical CatBoost framework that organizes substation cost prediction according to the engineering formation logic of project cost. The framework integrates intermediate-variable completion, target-wise cost prediction, sparse-target treatment, and engineering-constraint coordination. Compared with ordinary flat regression models, the proposed framework explicitly considers the layered relationship among basic engineering attributes, intermediate engineering quantities, final cost components, and aggregate investment indicators, thereby providing a more engineering-oriented modeling route.

Third, we introduce a structural coordination mechanism for multi-target cost prediction by incorporating predefined engineering accounting identities into the post-prediction stage. Instead of treating each target as an isolated regression output, the proposed method adjusts the predicted cost vector through reliability-weighted coordination, so that the final outputs better satisfy the balance relationships between component costs and total investment indicators. This design shifts the modeling objective from pure statistical fitting to a joint consideration of predictive accuracy, structural closure, and engineering feasibility.

Fourth, we provide a comprehensive empirical evaluation under realistic engineering data conditions. The proposed framework is compared with a strong single-stage CatBoost baseline and six conventional benchmark models, and is further analyzed through ablation experiments, trend-response visualization, feature-importance interpretation, and sparse-target discussion. The results reveal that the proposed framework outperforms conventional benchmark models and improves output consistency, while the strong CatBoost baseline remains competitive in some global statistical metrics. This finding clarifies the practical trade-off between statistical accuracy and engineering structural consistency in substation cost prediction.

The remainder of this paper is organized as follows: In Section 2, we describe the data source, preprocessing procedure, feature organization, and proposed hierarchical CatBoost framework. In Section 3, we present the experimental settings, evaluation metrics, and prediction results. In Section 4, we discuss ablation results, trend-level behavior, benchmark comparison, feature importance, limitations, and future work. In Section 5, we conclude the paper and summarize the practical implications of the proposed framework for substation project cost prediction.

2. Materials and methods

2.1. Data collection

The data used in this study were derived from engineering records compiled by the State Grid Economic and Technological Research Institute Co., Ltd. during project budget review and final account acceptance processes. The dataset mainly included the actual cost components of substation projects, together with a range of characteristic indicators reflecting project scale, technical schemes, construction conditions, and engineering attributes. Compared with simulated data or publicly available samples, these data were characterized by reliable provenance, clearly defined engineering attributes, and strong business relevance, thereby enabling a more realistic representation of the cost formation mechanisms of substation projects. We selected 2,820 State Grid substation project records in 2024 for analysis. After data preprocessing and organization, all 2820 records formed the structured dataset used for model training and validation, providing a solid data foundation for the subsequent cost prediction research.

2.2. Data cleaning and preprocessing

Data cleaning and preprocessing were conducted to improve training stability and predictive validity. In the raw engineering tables, abnormal expressions such as empty strings, “NA”, “NaN”, “null”, and “–” were uniformly treated as missing values. Blank or invalid historical records were checked and standardized according to the project identification information, thereby ensuring the consistency of the available engineering samples. After cleaning, the historical project records were organized into a structured sample set containing heterogeneous engineering features and multiple prediction targets for subsequent model development and evaluation. Subsequently, using a fixed random seed, the dataset was randomly partitioned into a training set containing 80% of the samples and an independent test set containing the remaining 20%. Model validation was conducted internally within the training set through the out-of-fold procedures described in the subsequent modeling stages.

Given the heterogeneous nature of the engineering records, a unified preprocessing strategy was adopted to convert mixed-type raw fields into structured model inputs. Geographic and scheme-related attributes were encoded as categorical features, voltage-level information was transformed into numerical magnitude features together with presence indicators, date fields were decomposed into interpretable temporal descriptors, boolean variables were converted into binary features, and continuous numerical fields were cleaned and standardized before numerical conversion. Through this procedure, the original multi-source engineering records were transformed into a consistent structured feature matrix suitable for subsequent hierarchical modeling.

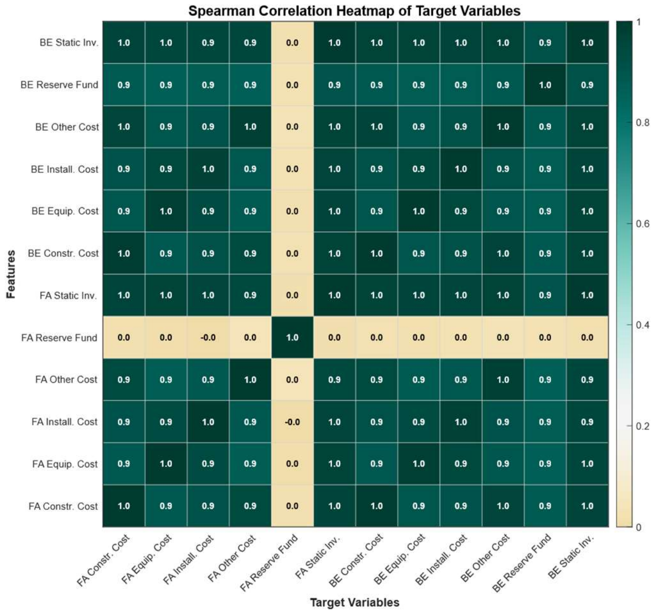
To match the hierarchical framework, the model variables were further organized into layered inputs and outputs. In the three-layer framework, the first-layer input set (Set A) contained 34 dimensions representing basic engineering attributes, whereas Set B contained 68 dimensions corresponding to intermediate variables that could be inferred from the basic attributes, resulting in a 102-dimensional feature representation for the subsequent target-prediction stage. In the second layer, the original basic inputs and the completed intermediate variables were jointly used to predict the target cost indicators. By contrast, the baseline model directly mapped the first-layer inputs to the target variables in a single-stage manner. All target cost variables were converted into numerical form before regression and subsequent structural coordination.

A unified variable-name mapping was used during model implementation to ensure consistent feature representation and result presentation.

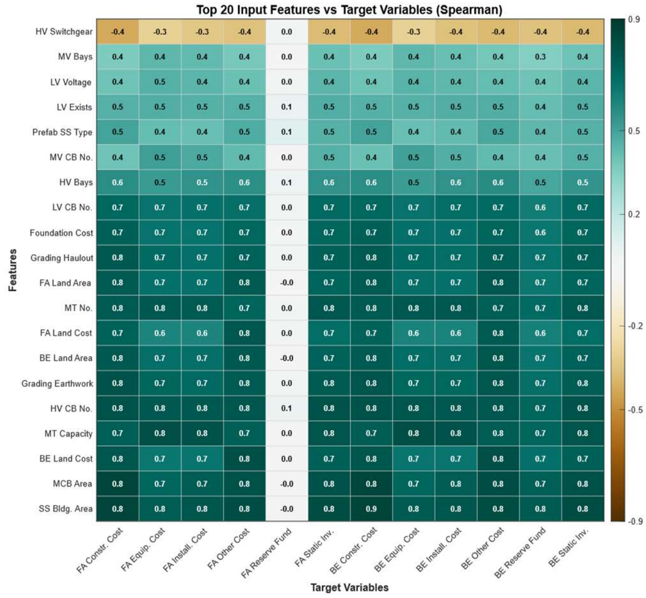
2.3. Feature organization and correlation analysis

To examine the correlation structure among different variables and support the hierarchical organization of the prediction task, Spearman correlation analysis was conducted at three levels: correlations among target variables, correlations between basic input variables and target variables, and correlations between intermediate variables and target variables. For the input and intermediate variables, the maximum absolute Spearman correlation coefficient with all targets was used as the screening criterion, and the top 20 variables were selected for visualization. Spearman correlation was adopted because the engineering cost data are characterized by skewed distributions, heterogeneous scales, zero values, and possible outliers, making rank-based correlation more suitable than linear correlation.

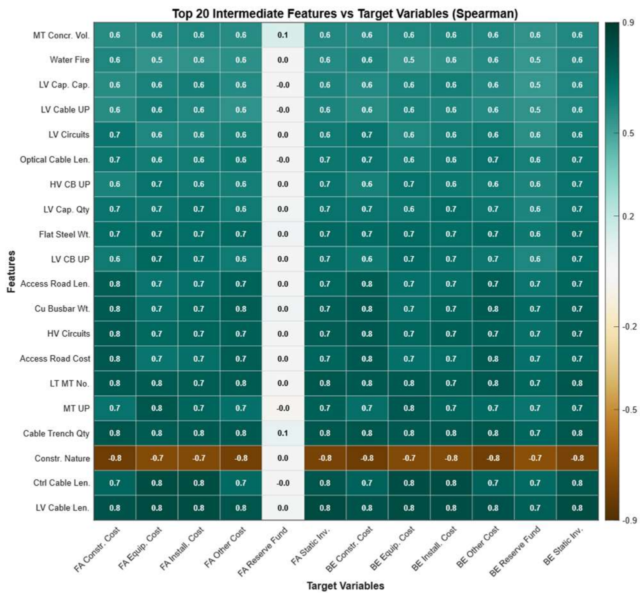
As shown in Figure 1, the correlation analysis reveals a clear hierarchical structure among target variables, basic input variables, and intermediate variables. Most major cost targets exhibit strong positive correlations, indicating that construction cost, equipment procurement cost, installation cost, other costs, and static investment are jointly affected by project scale and engineering configuration. The selected input variables with high correlations are mainly associated with building scale, land acquisition, voltage level, main transformer capacity, and equipment layout, whereas the highly correlated intermediate variables are more closely related to engineering quantities such as cable length, cable trench quantity, access road cost, and material weight. These results support the hierarchical organization of “basic inputs–intermediate variables–target variables”. In contrast, FA Reserve Fund shows consistently weak correlations with most variables, suggesting that it should be treated as a special sparse target rather than an ordinary cost component.



(a) Spearman correlation heatmap among target variables



(b) Spearman correlation heatmap between input variables and target variables



(c) Spearman correlation heatmap between intermediate variables and target variables

Figure 1. Spearman correlation heatmaps of target variables, input variables, and intermediate variables.

2.4. Overview of the model

We constructed a three-layer hierarchical CatBoost framework for multi-target cost prediction of substation projects, as illustrated in Figure 2. The core idea was to decompose the complex cost formation process into four connected stages, namely, basic feature representation, intermediate-variable completion, target prediction, and engineering-constraint coordination. Unlike single-stage direct regression, the proposed framework first reconstructed key intermediate variables from basic engineering attributes, then performed target-wise prediction based on the enriched representation, and finally coordinated the outputs through explicit engineering identities. In this way, the model aims to improve not only predictive capability, but also structural consistency and engineering feasibility.

The proposed framework was developed for cross-sectional engineering samples rather than strict rolling time-series forecasting. In other words, each project record was treated as an individual supervised-learning sample, and the model was trained to learn the mapping between engineering attributes and multiple cost targets under a structured tabular-data setting. This setting was consistent with the practical use scenario of project-level cost prediction in substation engineering.

After preprocessing, the heterogeneous engineering records were transformed into a structured feature space that served as the basis for the subsequent layered modeling procedure.

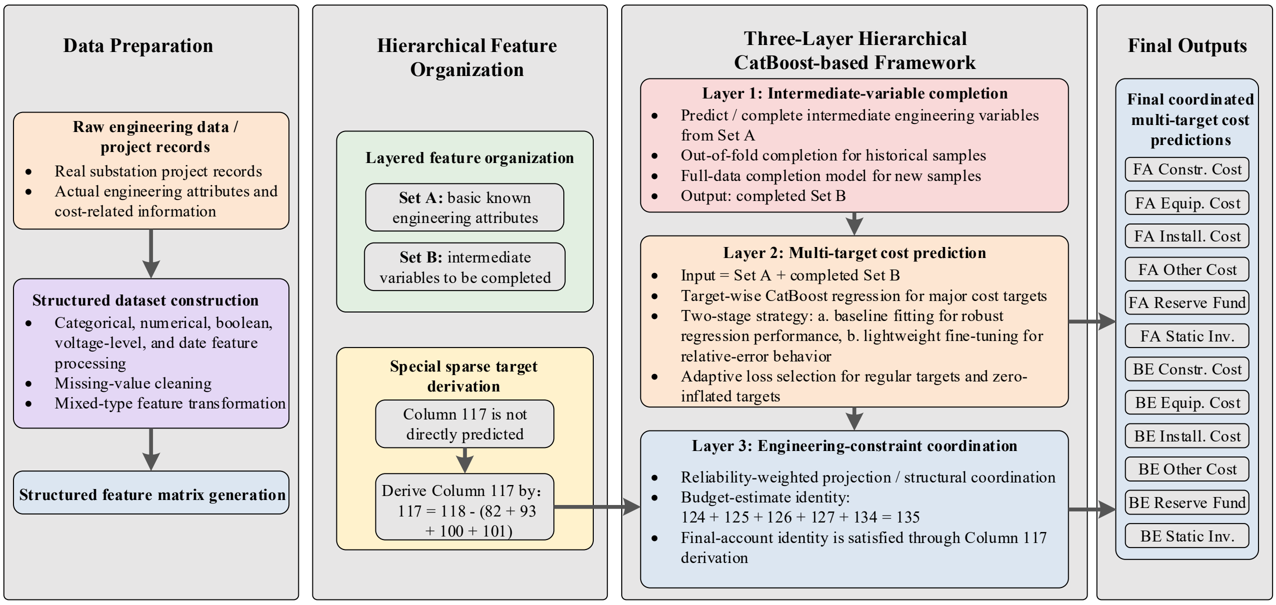


Figure 2. Overall framework of the proposed three-layer hierarchical CatBoost-based model for multi-target cost prediction of substation projects.

As shown in Figure 2, the framework consisted of data preparation, hierarchical feature organization, target-wise CatBoost prediction, sparse-target derivation, and engineering-constraint coordination. The following subsections describe the three key modeling stages in detail.

2.4.1. Intermediate-variable completion

In the first layer, the input variables were divided into two groups: Set A, which contained the basic known engineering attributes, and Set B, which contained intermediate variables to be completed from Set A. Our objective of this layer was not to predict cost directly, but to reconstruct a more informative intermediate representation from relatively stable basic project attributes. To reduce information leakage, out-of-fold completion was adopted within the training set, while the corresponding completion models fitted on the full training set were used to generate intermediate features for the independent test samples.

$$\hat{B} = f_1(A) \tag{1}$$

The completed Set B variables were then passed to the second layer together with the original Set A features.

2.4.2. Target-wise CatBoost prediction and sparse-target derivation

In the second layer, the completed intermediate variables from the first layer were concatenated with the original basic inputs to form the final feature representation for multi-target cost prediction. A target-wise CatBoost regression strategy was then adopted for the major cost components.

$$\widetilde{\mathbf{Y}}_i = f_2([\mathbf{A}_i, \widehat{\mathbf{B}}_i]) \tag{2}$$

To balance explanatory power and relative-error control, a two-stage training strategy was used. First, a baseline model was fitted to secure robust overall regression performance. Second, a lightweight fine-tuning step was introduced to improve relative-error behavior while avoiding substantial degradation in goodness of fit. The loss selection in the fine-tuning stage was made adaptively according to the target distribution. For targets with many zero or near-zero values, a more stable absolute-error-oriented loss was preferred, whereas for regular targets, a relative-error-oriented objective was used. This target-adaptive design was intended to balance fitting quality and relative-error control while reducing the instability caused by zero inflation and heterogeneous target scales.

The framework also introduced a structured derivation mechanism for special sparse targets. In particular, the final-account reserve fund, corresponding to Column 117 in the source data, was treated as a zero-inflated and highly volatile variable and was therefore derived from related subitems rather than modeled by a standalone regression task. The derivation followed the engineering identity 117 = 118 - (82 + 93 + 100 + 101), where Columns 82, 93, 100, 101, 117, and 118 corresponded to final-account construction cost, equipment procurement cost, installation cost, other cost, reserve fund, and static investment, respectively. For the i-th project sample, the second-layer CatBoost models first generated preliminary predictions for the regular cost targets. For the final-account side, the reserve fund was not treated as an ordinary learnable target because of its sparsity and residual-type accounting nature. Instead, after obtaining the predictions of the major final-account cost components and the aggregate static investment, it was derived as

$$\hat{y}_{i,117} = \hat{y}_{i,118} - (\hat{y}_{i,82} + \hat{y}_{i,93} + \hat{y}_{i,100} + \hat{y}_{i,101}). \tag{3}$$

In this way, the final-account accounting identity was satisfied by construction:

$$\hat{y}_{i,82} + \hat{y}_{i,93} + \hat{y}_{i,100} + \hat{y}_{i,101} + \hat{y}_{i,117} = \hat{y}_{i,118}. \tag{4}$$

This treatment was mainly intended to maintain structural closure and avoid producing independent predictions that violate the engineering identity. However, it did not necessarily guarantee improved target-wise accuracy for this sparse indicator.

2.4.3. Reliability-weighted engineering-constraint coordination

After the sparse final-account reserve fund was derived as described in Section 2.4.2, the third layer further coordinated the budget-estimate-side predictions through a projection strategy whose weights were defined by the inverse out-of-fold residual variances of the target models. The purpose of this layer was not to refit the cost targets, but to minimally adjust the preliminary budget-estimate predictions so that they satisfied the predefined accounting identity among budget-estimate cost components and static investment.

In the following formulation, source-data indices 124, 125, 126, 127, 134, and 135 corresponded to budget-estimate construction cost, equipment procurement cost, installation cost, other cost, reserve fund, and static investment, respectively. For the budget-estimate side, the preliminary prediction vector is defined as

$$\tilde{\mathbf{y}}_i^{BE} = [\tilde{y}_{i,124}, \tilde{y}_{i,125}, \tilde{y}_{i,126}, \tilde{y}_{i,127}, \tilde{y}_{i,134}, \tilde{y}_{i,135}]^{\mathrm{T}} \tag{5}$$

The corresponding accounting constraint is

$$\tilde{y}_{i,124} + \tilde{y}_{i,125} + \tilde{y}_{i,126} + \tilde{y}_{i,127} + \tilde{y}_{i,134} - \tilde{y}_{i,135} = 0 \tag{6}$$

Accordingly, the constraint matrix is written as

$$\mathbf{H}_{BE} = [1,1,1,1,1,-1] \tag{7}$$

The coordinated prediction vector is obtained by solving

$$\begin{gathered} \mathbf{z}_i^{BE} = \arg\min_{\mathbf{z}} \left(\mathbf{z} - \tilde{\mathbf{y}}_i^{BE}\right)^{\mathrm{T}} \mathbf{W}_{BE} \left(\mathbf{z} - \tilde{\mathbf{y}}_i^{BE}\right) \\ \text{s.t.} \quad \mathbf{H}_{BE}\mathbf{z} = 0 \end{gathered} \tag{8}$$

The reliability weight of each target was calculated according to the out-of-fold residual variance of the corresponding second-layer model. Since this variance reflects prediction errors on samples excluded from the corresponding model fitting, it was used as an empirical indicator of target-specific predictive reliability. Accordingly, its inverse was adopted so that more stable predictions received larger weights and undergo smaller adjustments during constraint coordination, as follows:

$$w_j = \frac{1}{\sigma_j^2 + \varepsilon}, \quad j \in \{124,125,126,127,134,135\}. \tag{9}$$

where σ_j^2 denotes the out-of-fold residual variance of the j-th target model and ε is a small positive constant used to avoid numerical instability. The weight matrix is defined as

$$\mathbf{W}_{BE} = \mathrm{diag}(w_{124}, w_{125}, w_{126}, w_{127}, w_{134}, w_{135}) \tag{10}$$

The closed-form solution is

$$\mathbf{z}_i^{BE} = \tilde{\mathbf{y}}_i^{BE} - \mathbf{W}_{BE}^{-1}\mathbf{H}_{BE}^{\mathrm{T}}\left(\mathbf{H}_{BE}\mathbf{W}_{BE}^{-1}\mathbf{H}_{BE}^{\mathrm{T}}\right)^{-1}\mathbf{H}_{BE}\tilde{\mathbf{y}}_i^{BE} \tag{11}$$

The final coordinated budget-estimate prediction is then set as

$$\hat{\mathbf{y}}_i^{BE} = \mathbf{z}_i^{BE} \tag{12}$$

To improve robustness and reproducibility, fixed random-control settings and conservative fallback strategies were adopted throughout the implementation. Model performance was subsequently evaluated at the target-specific and global levels using the metrics defined in Section 3.2.

In summary, the proposed method is a three-layer hierarchical CatBoost framework that integrates intermediate-variable completion, target-wise robust regression, special-target derivation, and engineering-constraint coordination. Its main contribution lies not in introducing a complex new deep-learning architecture, but in organizing multi-target substation cost prediction into a layered engineering-oriented process that jointly considers predictive performance, accounting-identity consistency, and practical applicability.

2.5. Confidentiality-preserving visualization strategy

Considering that the original engineering data and the prediction results involve sensitive cost information from actual projects, directly displaying the horizontal and vertical axes in their original scales could reveal key information such as project scale, site conditions, and investment level, which would be unfavorable for public publication and result presentation. Therefore, a confidentiality-preserving design was introduced during the plotting of the trend curves, in which the horizontal-axis variables and the vertical-axis predicted values were subjected to monotonic anonymization. For any original variable x, its median $\mathrm{med}(x)$ and interquartile range $\mathrm{IQR}(x)$ were first calculated based on the sample distribution, and robust standardization was then performed as follows:

$$z = \frac{x - \mathrm{med}(x)}{\mathrm{IQR}(x)} \tag{13}$$

On this basis, a monotonic compression based on the inverse hyperbolic sine function was applied to further suppress the influence of extreme values on the graphical scale. This procedure yielded the confidentiality-preserved display variable:

$$x^* = \mathrm{asinh}(z) \tag{14}$$

Because the transformation is monotonic, it preserves relative ordering and trend direction while reducing the risk of disclosing original cost scales.

In addition to the confidentiality-preserving treatment, the horizontal axes were further subjected to interpretable rescaling and smoothed visualization to improve the readability of the trend plots. The main reason is that, for altitude and building area, the sample distributions were markedly non-uniform in the original data. Many samples were concentrated in the lower range and relatively few samples were in the higher range. If the points were directly connected along the original horizontal axis, the resulting curves would tend to be overly crowded in the low-value region and excessively sparse in the high-value region, thereby weakening the recognizability of the overall trend. To address this issue, after completing the monotonic confidentiality-preserving transformation, we further organized the samples using a quantile-based binning strategy. Specifically, the horizontal-axis samples were divided into several intervals according to their quantiles, and the median predicted value of each cost item within each interval was taken as the representative point.

Let $\mathcal{I}_b$ denote the index set of samples belonging to the b-th quantile bin. The representative trend value of the j-th cost item in this bin is calculated as

$$\overline{y}_{b,j}^{\text{trend}} = \text{median}\left\{\hat{y}_{i,j}^{*} \mid i \in \mathcal{I}_b\right\}, \tag{15}$$

where $\hat{y}_{i,j}^{*}$ denotes the confidentiality-preserved predicted value after monotonic anonymization.

The median values obtained for each quantile bin were then connected sequentially to generate relatively smooth and interpretable trend curves. This treatment did not alter the sample ranking or the overall direction of variation. Instead, by aggregating information within local intervals, it reduced the interference of abnormal fluctuations from individual samples, thereby making the trend plots more suitable for presenting general patterns in academic papers.

3. Results

3.1. Experimental environment

The experiments in this study were conducted on a personal laptop, using a hybrid programming environment that integrates MATLAB and Python. Specifically, MATLAB was mainly employed for raw engineering data import, data cleaning and preprocessing, hierarchical feature construction, sample set partitioning, performance metric calculation, and result export. The Python environment was primarily used to invoke the CatBoost library for model training and prediction. In the implementation process, cross-environment interaction was achieved through the built-in Python interface in MATLAB, thereby enabling an effective integration of the data processing workflow with the regression model training procedure. To ensure the reproducibility of the experiments, fixed random seeds were set during sample partitioning and model training. This software environment provides adequate support for the construction, training, validation, and result analysis of the baseline CatBoost model and the proposed three-layer framework.

3.2. Evaluation metrics

To comprehensively evaluate the fitting capability and error performance of the constructed model in the task of substation project cost prediction, we selected the coefficient of determination (R^2), mean absolute percentage error (MAPE), symmetric mean absolute percentage error (sMAPE), and weighted mean absolute percentage error (wMAPE) as the core evaluation metrics. The selection considered the model implementation and the requirements of engineering applications. Among them, R^2 was mainly used to measure the explanatory power of the model with respect to the variation of the target variables, whereas MAPE, sMAPE, and wMAPE were primarily adopted to characterize the deviation between the predicted values and the true values from the perspective of relative error. Considering that different cost items in engineering cost data are consistent in dimension but may differ substantially in numerical scale, the use of relative-error-based metrics can more objectively reflect the overall performance of the model in a multi-target prediction scenario.

Let the total number of samples in the test set be denoted by N, the true value of the i-th sample by y_i, the predicted value by $\hat{y}_i$, and the mean of the true values by $\bar{y}$. Then, the evaluation metrics adopted in this study were defined as follows.

The coefficient of determination, R^2, was used to measure the explanatory power of the model with respect to the dispersion of the observed data, and it is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{16}$$

In general, the closer the R^2 value was to 1, the more adequately the model captured the variation pattern of the target variable, and the better the predictive fitting performance.

The mean absolute percentage error (MAPE) was used to measure the average deviation of the predicted values from the true values in relative terms, and it is defined as

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{17}$$

MAPE is readily interpretable, as it reflects the average prediction error in percentage form and thus facilitates horizontal comparison among different models. However, when the true value approaches zero, MAPE may become unstable or be excessively inflated. Therefore, for the prediction of certain small-value cost items in engineering cost estimation, the sole use of MAPE may be subject to limitations.

To mitigate the influence of small denominators on evaluation results, we further introduced the symmetric mean absolute percentage error (sMAPE), which is defined as

$$\text{sMAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\frac{2\,|\,y_i - \hat{y}_i\,|}{|\,y_i\,| + |\,\hat{y}_i\,| + \varepsilon} \tag{18}$$

where ε is a very small positive constant introduced to prevent the denominator from becoming zero. Compared with MAPE, sMAPE is more robust when the true value and the predicted value are small. To a certain extent, it can also mitigate the disturbance of extreme percentage errors on the overall evaluation results.

In addition, considering that different indicators in multi-target engineering cost prediction may vary substantially in monetary scale, we adopted the weighted mean absolute percentage error (wMAPE) to further characterize the overall prediction performance, which is defined as

$$\mathrm{wMAPE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i| + \varepsilon} \times 100\% \tag{19}$$

Compared with MAPE, wMAPE uses the total sum of the true values as the normalization basis, and therefore can better reflect the proportion of the overall error relative to the total scale. This metric is particularly suitable for prediction tasks such as engineering cost estimation, in which different cost items may vary substantially in magnitude.

Overall, R^2 measures the explanatory power of the model with respect to data variability from the perspective of goodness of fit, MAPE characterizes the average relative error level, sMAPE enhances the stability of evaluation under small-value target conditions, and wMAPE further reflects the comprehensive prediction deviation of the model from an overall weighted perspective under full-sample or multi-target settings. These metrics are complementary in terms of explanatory capability, local relative error, and overall relative error, and can therefore provide a relatively comprehensive evaluation of the predictive performance of the model constructed in this study.

It should be noted that, during the model development and internal analysis stages, the code in this study also computed additional metrics, including the mean absolute error (MAE), root mean square error (RMSE), and median absolute error (MedAE), to assist in assessing model performance and robustness on the absolute-error scale. However, given the sensitivity of engineering cost data, we mainly adopted R^2, MAPE, sMAPE, and wMAPE as the core publicly reported evaluation metrics in the formal presentation of results and model comparisons.

For the global metrics, all valid sample-target prediction pairs were aggregated across the evaluated target set. Let $\mathcal{J}$ denote the set of evaluated targets, and let

$$\Omega = \{(i,j) \mid i = 1,\ldots,N,\ j \in \mathcal{J}\} \tag{20}$$

denote the set of valid sample-target pairs. The global coefficient of determination is defined as

$$R^2_{\mathrm{global}} = 1 - \frac{\sum_{(i,j)\in\Omega} \left(y_{ij} - \hat{y}_{ij}\right)^2}{\sum_{(i,j)\in\Omega} \left(y_{ij} - \bar{y}_{\Omega}\right)^2}, \tag{21}$$

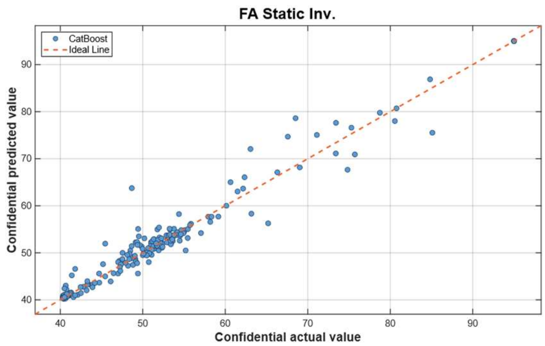
where $\bar{y}_{\Omega}$ is the mean of all observed values in Ω. The global sMAPE and global wMAPE are calculated as

$$\mathrm{sMAPE}_{\mathrm{global}}=\frac{100\%}{|\Omega|}\sum_{(i,j)\in\Omega}\frac{2\left|y_{ij}-\hat{y}_{ij}\right|}{\left|y_{ij}\right|+\left|\hat{y}_{ij}\right|+\varepsilon}, \tag{22}$$

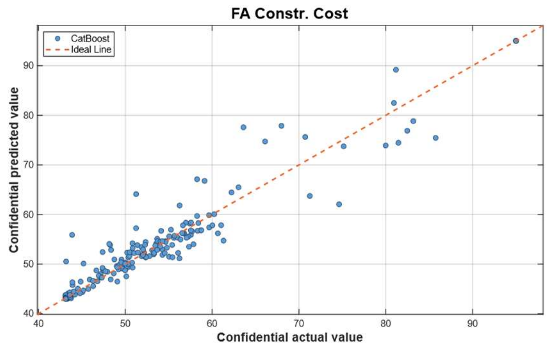
$$\mathrm{wMAPE}_{\mathrm{global}}=100\%\frac{\sum_{(i,j)\in\Omega}\left|y_{ij}-\hat{y}_{ij}\right|}{\sum_{(i,j)\in\Omega}\left|y_{ij}\right|+\varepsilon}. \tag{23}$$

3.3. Comparative analysis of parity plots under the baseline and optimized models

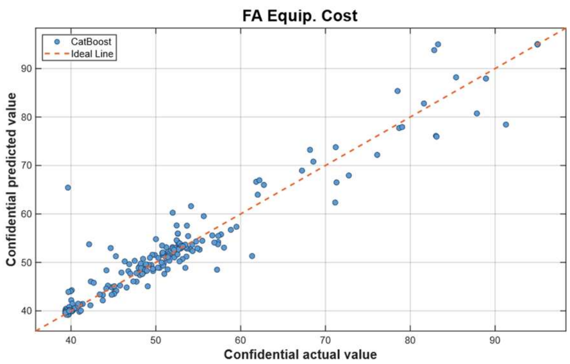
To visually compare the predictive behavior of the baseline CatBoost model and the optimized three-layer framework, parity plots between predicted and actual values were constructed for four representative final-account targets: construction cost, equipment procurement cost, installation cost, and static investment. To protect sensitive engineering cost information, the same confidentiality-preserving monotonic transformation was applied to actual and predicted values before plotting. Therefore, the figures mainly support visual comparison of relative fitting patterns, rather than disclosure of original monetary scales.



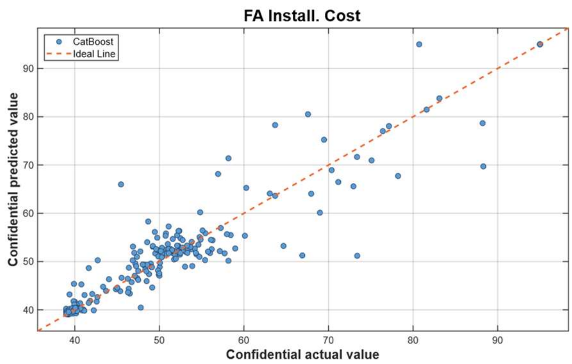
(a) Parity plot of final-account static investment under the baseline model



(b) Parity plot of final-account construction cost under the baseline model



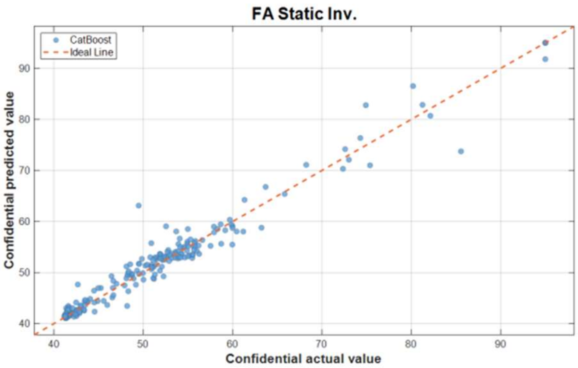
(c) Parity plot of final-account equipment procurement cost under the baseline model



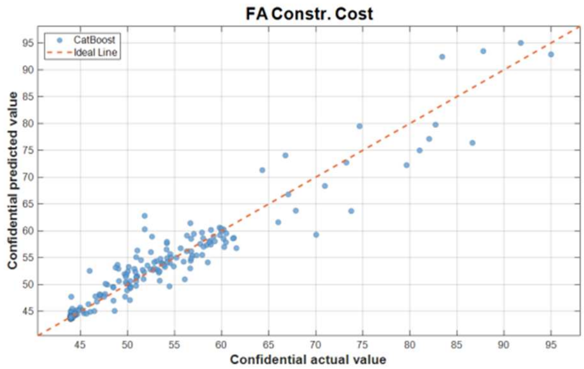
(d) Parity plot of final-account installation cost under the baseline model

Figure 3. Parity plots of representative targets under the baseline CatBoost model.

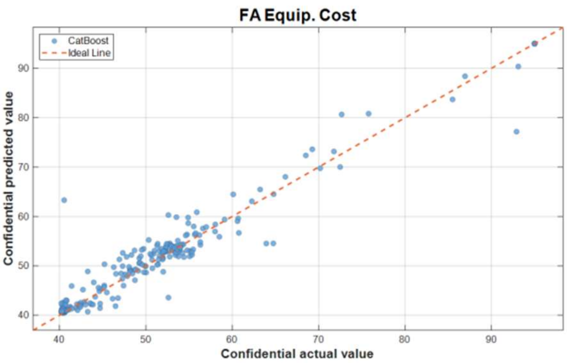
As shown in Figures 3 and 4, both models capture the overall increasing relationship between predicted and actual values. The optimized three-layer model shows slightly tighter local convergence for several targets, especially in heterogeneous value ranges, whereas the baseline CatBoost model remains highly competitive and performs better in some global statistical metrics, as reported in Tables 1 and 2. Therefore, the parity plots should be interpreted as descriptive evidence of local output coordination and engineering plausibility, rather than as proof of uniform statistical superiority.



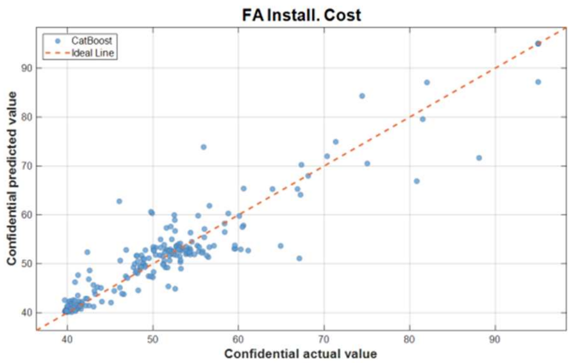
(a) Parity plot of final-account static investment under the optimized model



(b) Parity plot of final-account construction cost under the optimized model



(c) Parity plot of final-account equipment procurement cost under the optimized model



(d) Parity plot of final-account installation cost under the optimized model

Figure 4. Parity plots of representative targets under the optimized three-layer model.

3.4. Quantitative comparison between the baseline and optimized three-layer models

While we provide a visual comparison of the parity plots under the two modeling strategies in Section 3.3, the quantitative performance of the optimized three-layer model is summarized in Table 2. The table reports the target-wise and global results in terms of R^2, MAPE, sMAPE, and wMAPE, thereby enabling a more direct comparison with the baseline CatBoost model reported in Table 1. This comparison further evaluates the effectiveness of different modeling strategies in multi-target cost prediction.

A direct comparison between Tables 1 and 2 shows that the baseline CatBoost model achieves better global statistical performance than the optimized three-layer model on the dataset, with a higher Global R^2 (0.9671 vs. 0.9521) and a lower Global wMAPE (16.46% vs. 17.06%). This result indicates that the proposed hierarchical framework does not uniformly improve global error metrics over an already strong baseline. Nevertheless, the optimized model remains competitive on several major cost items and provides additional engineering-oriented benefits through structural coordination and the identity-based treatment of special sparse targets. Accordingly, its value should be interpreted primarily from the perspective of engineering feasibility and output consistency rather than simple metric dominance.

The calculation of global metrics follows the aggregation strategy defined in Section 3.2.

Before interpreting the target-wise results, it is necessary to distinguish regular cost targets and the special sparse accounting target. In this study, most cost indicators, including construction cost, equipment procurement cost, installation cost, other cost, and static investment, were treated as regular prediction targets because they are directly associated with project scale, equipment configuration, civil engineering quantities, and construction conditions. By contrast, the final-account reserve fund is treated as a special sparse accounting target because it is derived from the residual relationship among final-account cost components and the aggregate investment indicator.

Therefore, the target-wise result of FA Reserve Fund is retained in the table for transparency, but it should not be interpreted in the same way as the regular cost components. To avoid misleading conclusions, the overall performance of the proposed framework is discussed from two perspectives: All evaluated targets and regular targets excluding the final-account reserve fund.

Table 1. Experimental results of the baseline model.

Target	R^2	MAPE	sMAPE	wMAPE
FA Constr. Cost	0.9175	914.40	74.66	20.55
FA Equip. Cost	0.9464	190.04	44.35	17.50
FA Install. Cost	0.8639	84.43	40.47	24.26
FA Other Cost	0.9085	104.41	44.61	19.43
FA Reserve Fund	0.4626	69.49	198.11	110.43
FA Static Inv.	0.9636	88.25	37.15	14.04
BE Constr. Cost	0.9361	198.32	73.07	18.07
BE Equip. Cost	0.9560	147.96	43.56	16.90
BE Install. Cost	0.8654	172.22	36.21	21.36
BE Other Cost	0.8991	104.94	43.73	18.83
BE Reserve Fund	0.8193	77.99	51.47	23.72
BE Static Inv.	0.9663	69.31	35.32	14.03
Global Metrics	0.9671	-	60.23	16.46

The result of the FA Reserve Fund requires separate interpretation. As shown in Table 2, this target exhibits a strongly negative R^2 and extremely large percentage-based errors. This result does not indicate a general failure of the proposed framework on multi-target cost prediction. Rather, it reflects the special statistical and accounting nature of this indicator. In this dataset, the final-account reserve fund behaves as a sparse residual-type variable rather than as a regular engineering-scale-driven cost component. Its true values are often zero or very small, and its magnitude is determined by the residual relationship between aggregate investment and several major component costs. Consequently, even a small absolute deviation in the predicted aggregate or component costs can be amplified into a very large relative error for this target.

Table 2. Predictive performance of the proposed hierarchical model for regular and special cost targets.

	Target	R^2	MAPE	sMAPE	wMAPE
Panel A. Regular cost targets	FA Constr. Cost	0.9278	178.39	67.00	20.13
	FA Equip. Cost	0.9391	133.77	38.21	16.63
	FA Install. Cost	0.8988	102.07	35.04	21.93
	FA Other Cost	0.8663	104.98	40.65	21.29
	FA Static Inv.	0.9629	59.27	31.07	12.89
	BE Constr. Cost	0.8844	128.38	63.62	20.73
	BE Equip. Cost	0.9529	74.07	36.12	15.49
	BE Install. Cost	0.7804	86.99	33.94	24.13
	BE Other Cost	0.8987	86.14	37.50	19.56
	BE Reserve Fund	0.9064	65.47	47.17	20.79
	BE Static Inv.	0.9639	62.99	31.20	13.41
	Global Metrics	0.9521	-	54.89	17.06
Panel B. Special sparse accounting target	FA Reserve Fund	-61.9648	2776.05	197.17	3596.18

* Note: FA Reserve Fund is reported as a diagnostic sparse accounting target and is not included in the main global metrics used for model comparison.

For this reason, the FA Reserve Fund is retained in the results for transparency but is not used as a representative target for assessing the general predictive capability of the model. The main performance interpretation should focus on the regular cost components and aggregate indicators, including construction cost, equipment procurement cost, installation cost, other cost, and static investment. From this perspective, the proposed framework maintains competitive prediction accuracy while providing additional structural consistency through engineering-constraint coordination.

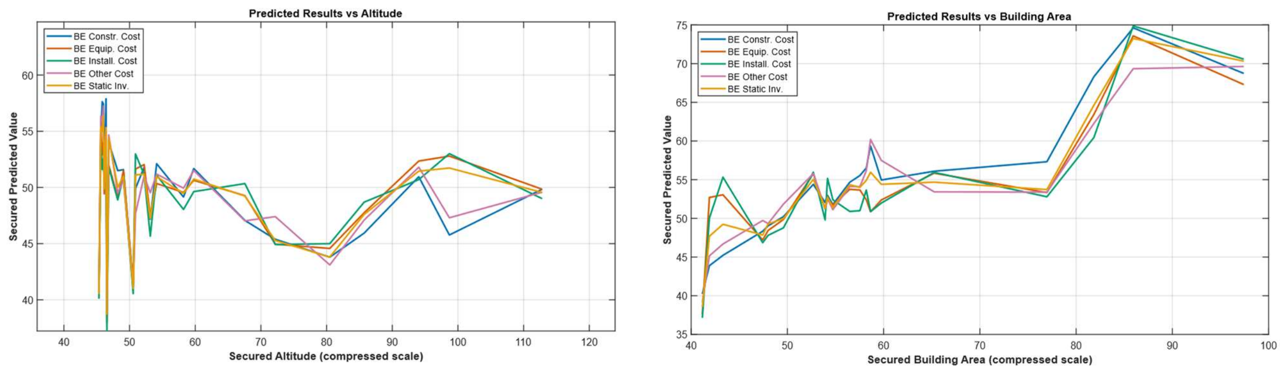
3.5. Trend response analysis of multi-target cost predictions

To further reveal the response patterns of the optimized model to multiple cost items under different engineering conditions, we, based on the analysis of prediction results, plotted trend curves with project altitude and building area as the horizontal axes and the predicted values of five major budget-estimate cost items excluding the reserve fund as the vertical axes. Unlike simple scatter plots comparing predicted and actual values, such trend plots place greater emphasis on the overall evolution patterns of model outputs as typical engineering conditions vary continuously. Therefore, from the perspective of continuous variation, they make it possible to reflect the sensitivity of different cost components to key engineering attributes as well as their coordinated variation relationships. Accordingly, our primary purpose of this analysis was not to examine prediction errors point by point but rather to identify, at a global level, the trend characteristics of cost items as altitude and building area change. This analysis thereby provides a visual basis for subsequent discussion of engineering mechanisms and economic interpretation.

Considering the confidentiality requirements of real engineering cost data, the trend plots were generated using the confidentiality-preserving visualization strategy described in Section 2.5. Therefore, the axes in Figure 5 should be interpreted as anonymized monotonic display scales rather than original engineering or monetary values. This treatment preserves the relative trend patterns while avoiding direct disclosure of sensitive project information.

Figure 5(a) shows that the five cost curves fluctuate with altitude and do not follow a strictly monotonic pattern. This indicates that altitude is more likely to influence substation cost indirectly through site conditions, transportation difficulty, construction organization, and foundation treatment, rather than acting as a single dominant cost driver. Despite local fluctuations, the relative hierarchy among static investment, equipment procurement cost, construction cost, installation cost, and other costs remains generally reasonable, suggesting that the model preserves coordinated cost relationships under changing environmental conditions.

By contrast, Figure 5(b) presents a clearer upward trend with building area. This is consistent with engineering practice, because building area directly reflects project scale, spatial configuration, civil workload, and supporting facility requirements. Overall, the two trend plots suggest that the optimized model can capture different types of engineering drivers: Altitude mainly represents external construction conditions, whereas building area reflects internal scale expansion.



(a) Trend curves of five estimated cost items with respect to project altitude

(b) Trend curves of five estimated cost items with respect to building area

Figure 5. Trend curves of five estimated cost items under the optimized model with respect to project altitude and building area.

4. Discussion

To address the challenges of strong multi-target coupling, high data heterogeneity, and difficult error control in substation project cost prediction, we constructed a proposed baseline model and a three-layer optimized model and validated them using actual engineering data. The results show that the baseline CatBoost model is already highly competitive in terms of global statistical metrics, whereas the proposed three-layer framework provides a more structured engineering-oriented modeling framework. Accordingly, the practical value of the proposed method lies not in outperforming the baseline on all global accuracy indicators, but in improving output feasibility, structural closure, and the treatment of selected sparse or derived targets under engineering constraints.

In this section, we discuss the implications of the empirical results from the perspectives of module contribution, engineering consistency, target-specific behavior, and practical applicability. Rather than repeating the numerical comparisons reported in Section 3, for the discussion, we focus on what the results imply for hierarchical modeling in substation project cost prediction, especially under the coexistence of strong baseline performance, heterogeneous target behavior, and engineering-constraint requirements.

4.1. Ablation study

To verify the contribution of each component in the proposed three-layer CatBoost framework and to reveal the effects of different modules on prediction accuracy and engineering consistency, six ablation schemes were designed in this study. The six schemes were as follows: The baseline model that directly predicts the target variable set Y using only the basic input variable set A (Baseline CatBoost); the theoretical upper-bound model that jointly uses A and the true intermediate variables $B^{(\text{true})}$ to predict Y ($A+B(\text{True})$); a two-layer model that employs the first layer to predict the intermediate variables $\hat{B}$ and then performs target prediction in the second layer without imposing structural constraints (Two-layer w/o Constraint); a model that adds third-layer constraint coordination to the above two-layer model while directly predicting the target in the reserve fund (Two-layer + Constraint); the complete three-layer model that further adopts the derivation strategy for the reserve fund (Full Three-layer Model); and a variant that retains the full three-layer structure but removes the subsequent fine-tuning process in the second layer, preserving only the initial RMSE-based training stage (Full Model w/o Fine-tuning).

For the schemes incorporating structural constraints, coordinated correctionwas is performed at the prediction-output level for selected targets so that the results better satisfy the predefined engineering identities. To evaluate the structural consistency of each ablation scheme, the residuals of the two accounting identities were calculated for each sample as

$$r_{1,i} = \hat{y}_{i,82} + \hat{y}_{i,93} + \hat{y}_{i,100} + \hat{y}_{i,101} + \hat{y}_{i,117} - \hat{y}_{i,118}, \tag{24}$$

$$r_{2,i} = \hat{y}_{i,124} + \hat{y}_{i,125} + \hat{y}_{i,126} + \hat{y}_{i,127} + \hat{y}_{i,134} - \hat{y}_{i,135}. \tag{25}$$

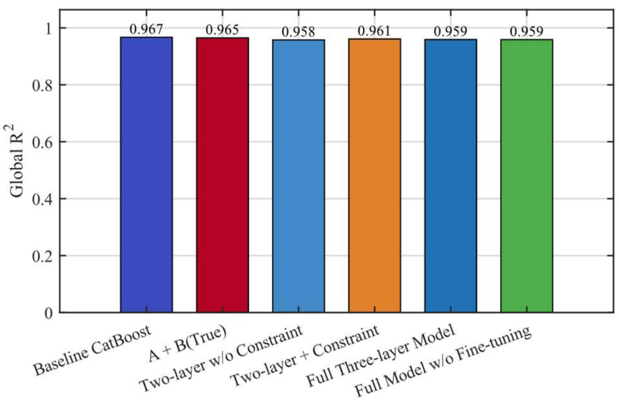
The mean absolute constraint residual is then defined as

$$C_k = \frac{1}{N}\sum_{i=1}^{N}|r_{k,i}|, \quad k = 1, 2. \tag{26}$$

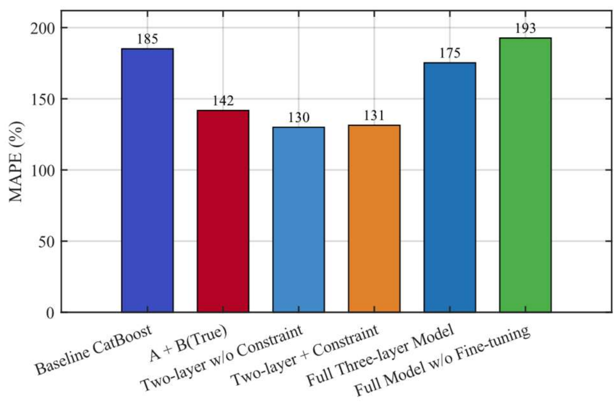
As shown in Figures 6–8, the ablation results reveal a trade-off between statistical accuracy and engineering consistency. The baseline CatBoost model achieves the best global R^2 and RMSE, indicating that the basic input variables alone already provide a strong predictive basis. The A+B (True) scheme performs slightly better in MAE and wMAPE, suggesting that the intermediate variables contain useful information when they are accurately available. However, replacing true intermediate variables with predicted ones introduces error propagation, which explains the performance decline of the unconstrained two-layer model and indicates that the effectiveness of hierarchical feature enrichment depends on the reliability of intermediate-variable completion.

The introduction of third-layer constraint coordination substantially improves structural consistency by reducing the residuals of the engineering balance constraints to zero or near zero.

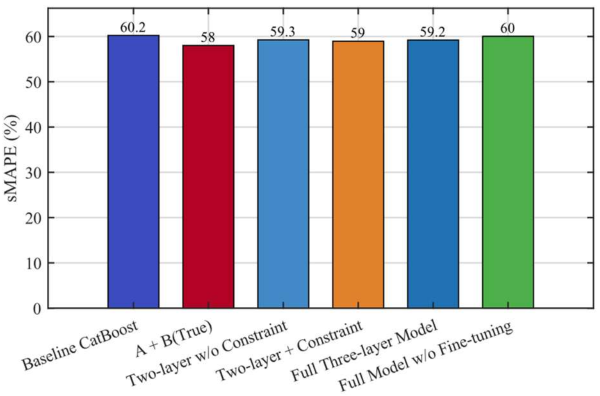
Although the full three-layer model does not uniformly outperform the strong single-stage CatBoost baseline in global statistical metrics, this difference can be attributed to error propagation from the predicted intermediate variables and the subsequent adjustment of preliminary predictions required to enforce the accounting identities. The resulting framework therefore provides more structurally consistent and engineering-feasible outputs rather than unconditional gains in predictive accuracy. In budget-review applications, consistency between disaggregated cost components and aggregate investment is essential. The ablation results therefore indicate that the observed limited loss in statistical accuracy represents a reasonable trade-off for improved structural closure and output feasibility.



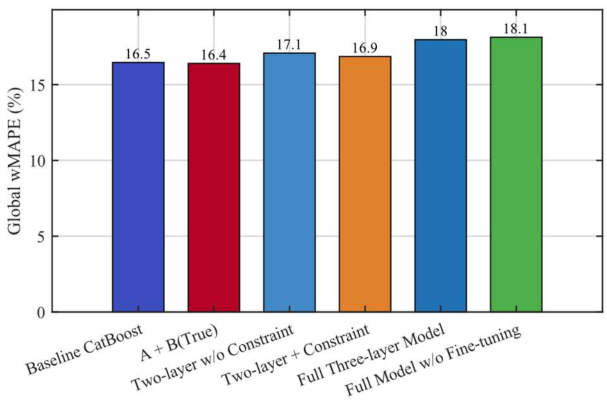
(a) Global R^2



(b) MAPE

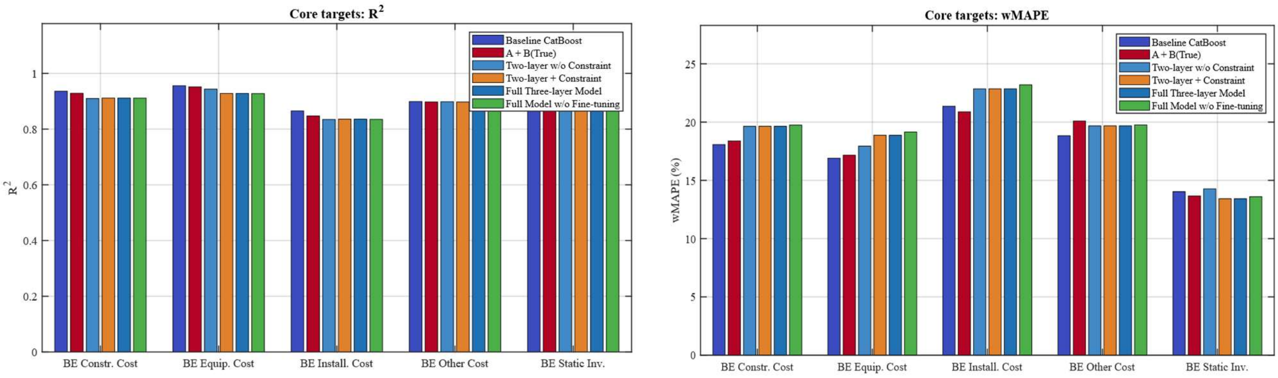


(c) sMAPE



(d) Global wMAPE

Figure 6. Global ablation results.



(a) Target-specific R^2 for core cost indicators

(b) Target-specific wMAPE for core cost indicators

Figure 7. Target-specific ablation results for core cost indicators.

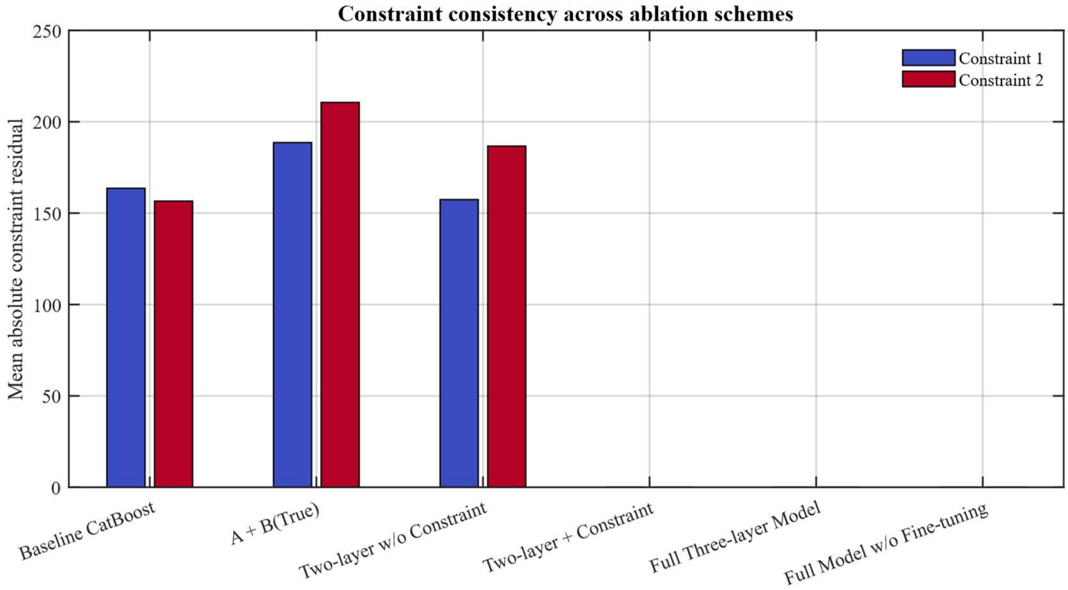
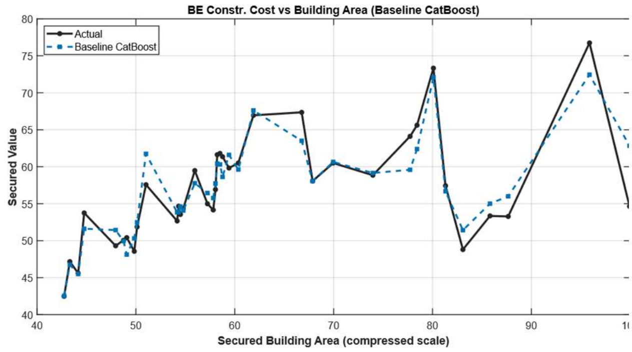


Figure 8. Constraint residuals.

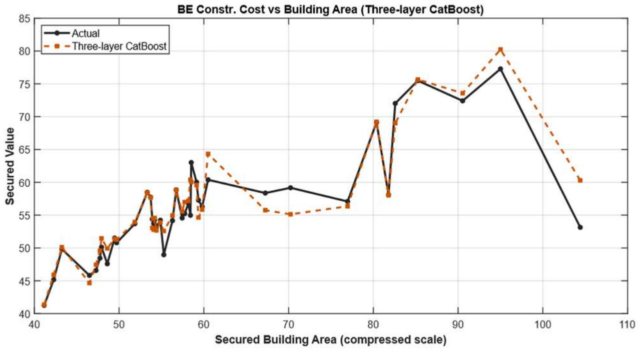
4.2. Trend-level comparison under building-area variation

To compare the trend-level behavior of the baseline CatBoost model and the optimized three-layer model under project-scale variation, predicted and actual trend curves were plotted for five budget-estimate cost items using building area as the horizontal-axis variable. The curves were generated using confidentiality-preserving transformation and quantile-based median aggregation, so that the dominant trend patterns could be displayed without exposing original project-scale or cost information.

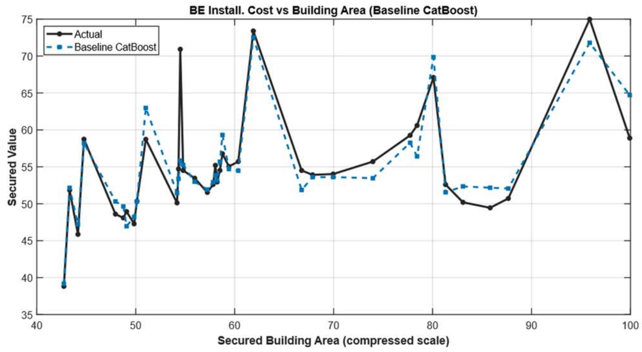
As shown in Figure 9, both models capture the overall upward relationship between building area and the five cost items, indicating that building area is an important scale-related driver of substation project cost. Compared with the baseline model, the optimized three-layer model shows closer visual agreement with the actual curves in several medium-to-high building-area intervals and near local turning points. This result suggests that the hierarchical use of intermediate variables and structural coordination can improve trend-level interpretability and local output consistency, although this visual advantage should be interpreted with the global quantitative metrics reported in Tables 1 and 2.



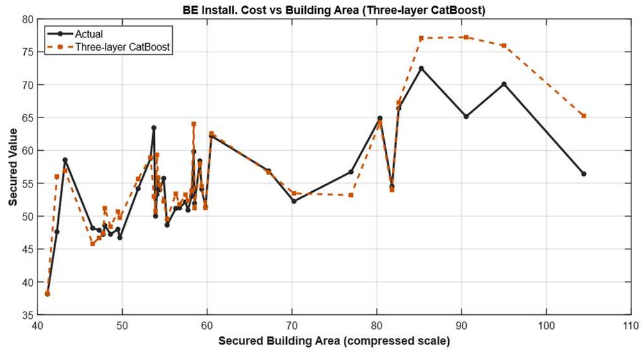
(a) Baseline model: construction cost



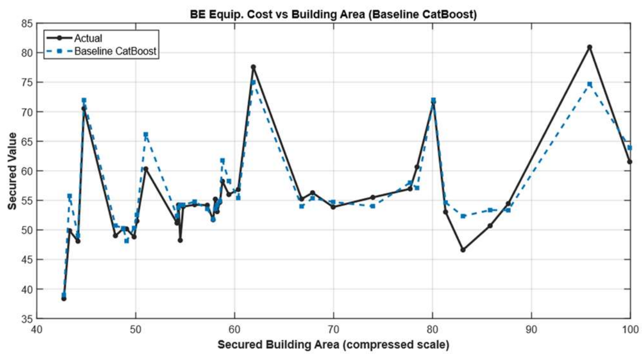
(b) Three-layer model: construction cost



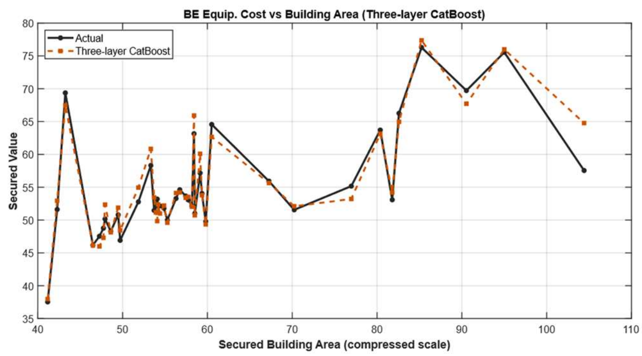
(c) Baseline model: installation cost



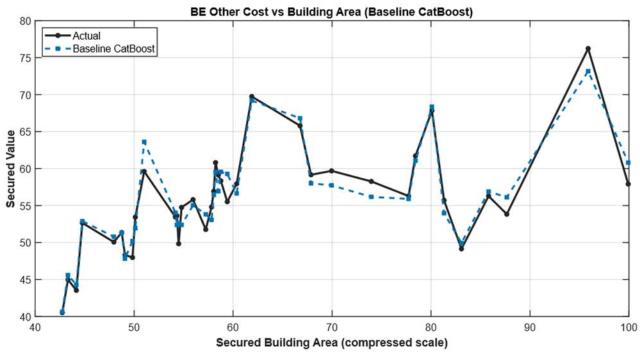
(d) Three-layer model: installation cost



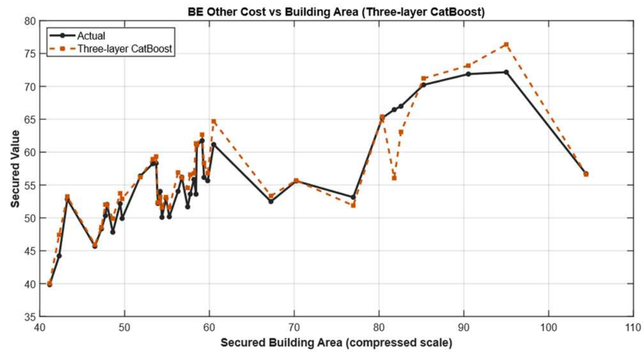
(e) Baseline model: equipment procurement cost



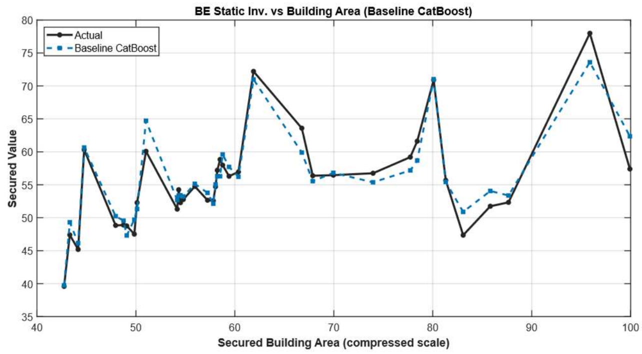
(f) Three-layer model: equipment procurement cost



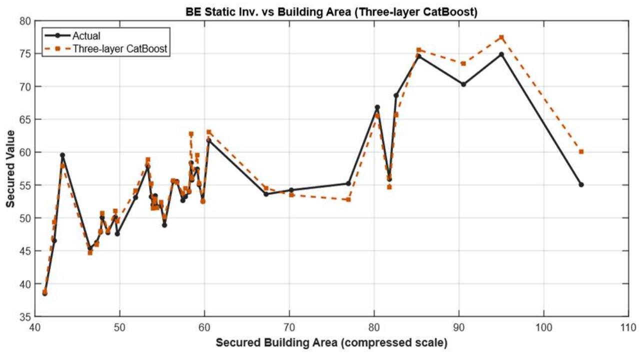
(g) Baseline model: other costs



(h) Three-layer model: other costs



(i) Baseline model: total cost



(j) Three-layer model: total cost

Figure 9. Comparison between actual and predicted values of the baseline CatBoost model and the optimized three-layer model for five budget-estimate cost items under building-area variation.

Figure 10 further compares the training and validation RMSE convergence curves of the baseline CatBoost model and the optimized three-layer model for budget-estimate static investment, corresponding to Column 135 in the source data, as a representative aggregate target. Both models show rapid error reduction in the early iterations and gradual stabilization in the later stage, indicating stable boosting behavior without obvious divergence. The three-layer model achieves lower training error and comparable validation behavior for this representative target, suggesting stronger fitting capacity after intermediate-variable completion. However, this convergence result should not be overinterpreted as global statistical dominance, but rather as supplementary evidence for the numerical stability and feasibility of the hierarchical modeling design.

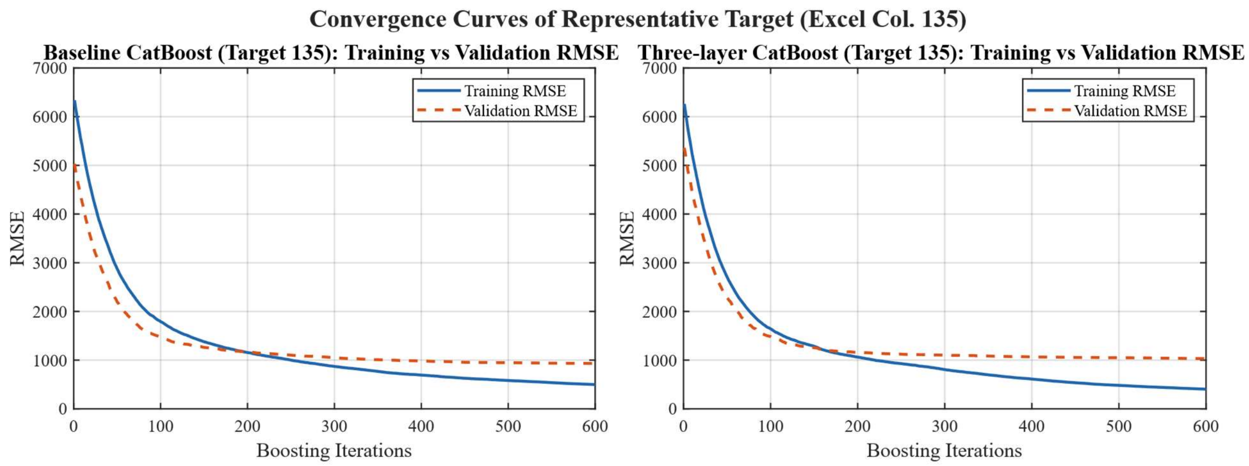


Figure 10. Iteration curves of the baseline model and the optimized model.

4.3. Comparison with conventional benchmark models

Table 3 compares the proposed framework with six conventional benchmark models. The results show that tree-based ensemble models, including Bagged Trees and LSBoost Trees, outperform Regression Tree, SVM, KNN, and Ridge Linear, indicating that nonlinear ensemble methods are more suitable for the heterogeneous engineering features used in this study.

Compared with these conventional benchmark models, the proposed three-layer framework achieves the best global R^2 and global wMAPE. This demonstrates that hierarchical feature organization and CatBoost-based target-wise prediction provide stronger predictive capability than ordinary machine-learning baselines under the current substation cost dataset.

However, this comparison should be distinguished from the comparison with the strong single-stage CatBoost baseline in Table 1. The proposed framework does not uniformly dominate the strong CatBoost baseline in all statistical metrics; its main advantage lies in structural consistency, accounting-feasible outputs, and the explicit treatment of sparse residual-type targets.

Table 3. Performance of six additional benchmark models on the current dataset.

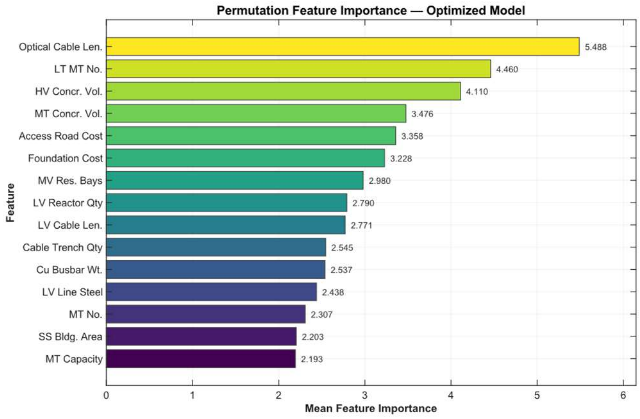
Model	Regression Tree	Bagged Trees	LSBoost Trees	SVM (RBF)	KNN	Ridge Linear
Mean_ R^2	0.7526	0.8478	0.8674	0.1186	-0.0143	-0.0143
Median_ R^2	0.8065	0.8914	0.8831	0.0795	-0.0152	-0.0152
Mean_MAPE	151.41	222.00	98.00	1283.11	4512.41	4512.41
Mean_sMAPE	52.61	60.71	54.65	87.64	100.72	100.72
Mean_wMAPE	37.14	30.61	29.30	59.87	91.99	91.99
Global_ R^2	0.8336	0.9363	0.9360	0.0647	-0.0167	-0.0167
Global_sMAPE	52.61	60.71	54.65	87.64	100.72	100.72
Global_wMAPE	24.95	19.20	17.97	61.47	84.28	84.28

4.4. Feature importance analysis

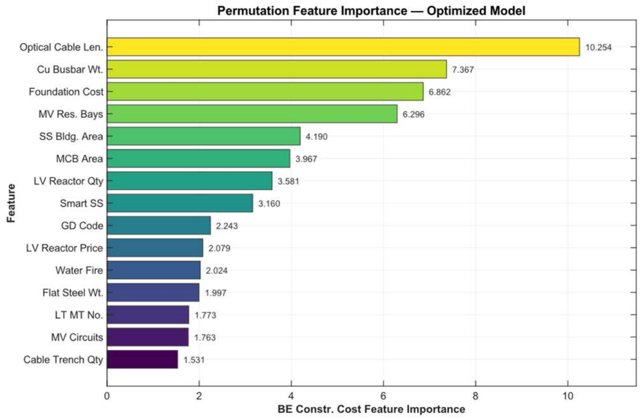
To further reveal the decision basis of the proposed hierarchical prediction framework and to identify the key driving factors affecting different cost targets, we conducted a feature importance analysis for the second-layer CatBoost prediction models. Specifically, after the training of each target variable was completed, permutation-based feature importance scores were computed for the second-layer CatBoost prediction models and ranked from the comprehensive perspective and the individual-target perspective. For the comprehensive feature importance analysis, the feature importance results of all valid target models were averaged, and the top 15 variables ranked by average importance were selected for presentation, as shown in Figure 11(a). For the five key targets, namely, estimated construction cost, estimated equipment procurement cost, estimated installation cost, estimated other costs, and estimated static investment, the top 15 important features of their respective models were extracted, and the results are shown in Figure 11(b)–11(f). This analysis reflects the dependence of the model on input variables from global and local perspectives, thereby enhancing the interpretability of the prediction framework. This design is consistent with studies emphasizing that applied machine-learning models should be evaluated not only by prediction accuracy, but also by their ability to provide interpretable evidence for engineering decision-making [13,27,32].

The comprehensive feature-importance results show that optical cable length, transformer planning variables, concrete quantities, access road cost, foundation treatment cost, cable length, and building area are among the most influential features. These variables cover equipment scale, civil engineering quantity, site accessibility, material configuration, and building scale, indicating that the model captures multiple dimensions of substation cost formation rather than relying on a single dominant factor.

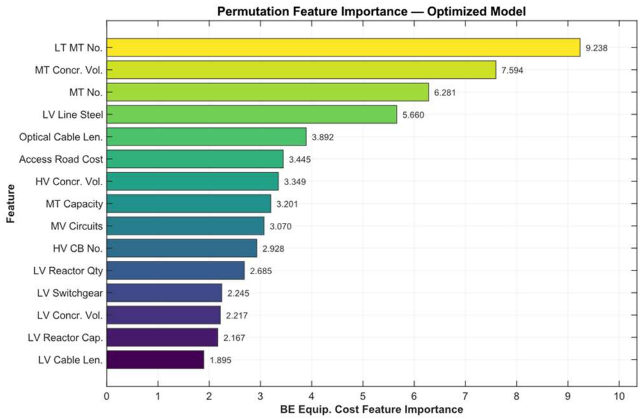
The target-specific results further show both shared and differentiated cost drivers. Transformer- and cable-related variables frequently appear across targets, while building-area variables contribute more strongly to construction cost, material and cable variables are more relevant to installation cost, and site-development variables are more important for other costs. This pattern enhances the interpretability of the proposed framework and supports its engineering applicability.



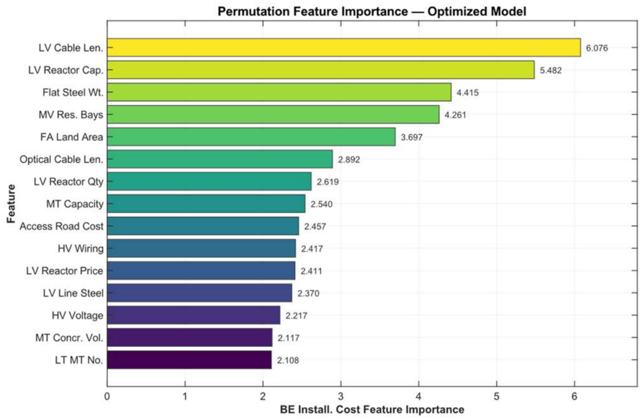
(a) Comprehensive feature importance analysis of the optimized model



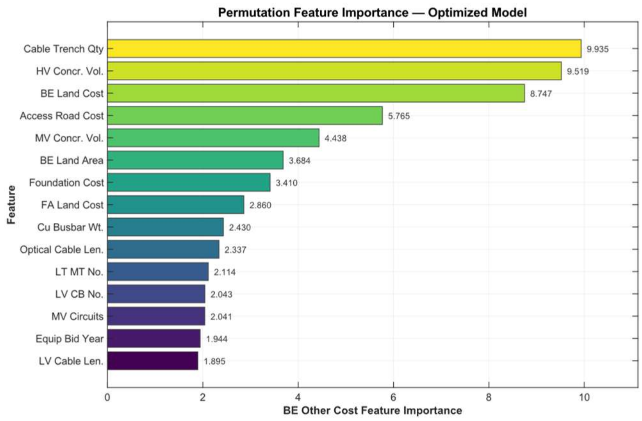
(b) Feature importance analysis of estimated construction cost



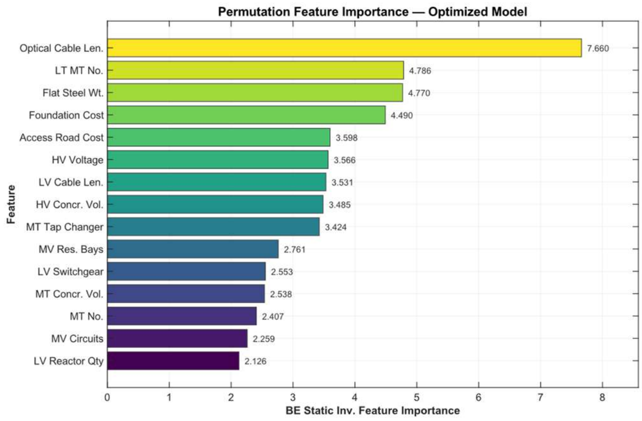
(c) Feature importance analysis of estimated equipment procurement cost



(d) Feature importance analysis of estimated installation cost



(e) Feature importance analysis of estimated other costs



(f) Feature importance analysis of estimated static investment

Figure 11. Feature importance analysis of the hierarchical prediction model.

4.5. Interpretation of the special sparse accounting target

The final-account reserve fund behaves differently from regular cost components. It is closely related to accounting adjustment, project-specific reserve use, and the residual relationship between aggregate investment and component costs, rather than being mainly determined by ordinary engineering-scale variables. Therefore, its weak target-wise performance should be interpreted as evidence of the limitations of standard regression when applied to sparse residual-type accounting targets, rather than as a failure of the overall multi-target prediction framework.

Accordingly, the weak predictive performance of the final-account reserve fund reflects its sparse, residual-type accounting nature rather than a limitation of the overall framework; its identity-based derivation is intended primarily to ensure structural closure and does not improve target-specific accuracy. In future work, researchers may therefore consider two-stage classification–regression, zero-inflated, quantile-regression, or uncertainty-aware approaches to better model this indicator while retaining the hierarchical multi-target prediction and engineering-constraint coordination structure.

4.6. Limitations and future work

Although the baseline model and the optimized three-layer model constructed in this study achieved favorable results in the task of substation project cost prediction and demonstrated strong overall advantages over multiple comparative models, several limitations remain, which also point to promising directions for further optimization of the applied machine-learning framework.

1. The data used in this study were derived from actual engineering records generated during budget review and final-account acceptance processes and therefore possess strong authenticity and engineering representativeness. However, this sample scope is mainly concentrated on substation projects within the State Grid system in 2024, and the generalization capability of the framework across years, regions, voltage levels, construction types, and other power-infrastructure project categories requires further validation. Researchers may therefore incorporate more diverse engineering samples and, where necessary, adapt project-specific feature and accounting definitions to improve the robustness and broader applicability of the framework.
2. The proposed model mainly relies on structured features obtained from engineering cost review and final-account data, including project scale, technical conditions, regional attributes, and related cost indicators. However, external factors such as material-price fluctuations, regional labor-cost differences, construction periods, policy adjustments, and equipment-market conditions have not been systematically incorporated. Researchers could build upon the structured variables by further integrating external information such as macroeconomic indicators, regional market conditions, and construction-environment factors, thereby establishing a multi-source data fusion framework to enhance the model's adaptability to complex engineering scenarios.
3. The proposed model achieved relatively stable predictive performance for most major cost indicators, whereas some reserve-related or small-value indicators were affected by zero values, sparse distributions, and strong fluctuations. Researchers may adopt classification – regression joint modeling, zero-inflated learning, quantile regression, or uncertainty-aware prediction to improve the identification and prediction of these special accounting targets

while retaining the hierarchical multi-target prediction and engineering-constraint coordination structure.

Taken together, these observations clarify both the practical value and the limitations of the proposed hierarchical framework, which are summarized in the following conclusion section. In particular, construction-material price fluctuations may introduce temporal shifts into cost data, and dedicated price-prediction or market-index variables could be incorporated in future extensions [33].

5. Conclusions

In this study, we proposed a constraint-coordinated hierarchical CatBoost framework for multi-target cost prediction in substation engineering projects. Using 2820 real engineering records collected within the State Grid system in 2024, the proposed framework decomposes the prediction task into basic feature representation, intermediate-variable completion, target-wise cost prediction, and engineering-constraint coordination. This design reflects the practical formation logic of substation project cost, in which multiple cost components are not independent labels but are associated with engineering quantities, project-scale attributes, and accounting identities. The principal findings are summarized as follows:

- The experimental results show that the proposed framework achieves a global R^2 of 0.9521 and a global wMAPE of 17.06%, demonstrating strong predictive capability on real-world heterogeneous engineering data. Compared with six conventional benchmark models, the proposed framework achieves better global fitting performance and overall weighted error control. These results indicate that hierarchical feature organization and CatBoost-based target-wise prediction can effectively support cost estimation for complex substation projects.
- The comparison with the strong single-stage CatBoost baseline shows that the proposed framework does not achieve the best result for every global statistical metric, and the baseline remains competitive in overall predictive accuracy. The principal contribution of the proposed framework lies in coordinating multi-target outputs: it reduces violations of predefined engineering accounting identities and improves consistency between disaggregated cost components and aggregate investment indicators. The resulting framework therefore provides a practical balance between statistical accuracy and accounting-identity consistency while demonstrating the value of task-specific hierarchical organization and constraint-based output optimization in applied machine-learning prediction.
- The target-wise analysis indicates relatively stable performance for the major cost components, whereas reserve-related indicators require separate treatment because of their sparse, zero-inflated, and residual-type characteristics. Target-adaptive strategies, including classification-regression modeling, zero-inflated learning, quantile prediction, and uncertainty-aware estimation, may provide more suitable mechanisms for these special accounting targets.

Overall, the framework can support budget review, investment control, and engineering decision-making through structurally consistent multi-target cost estimates. In future research, we will extend the dataset across years, regions, voltage levels, and project types, integrate external market and construction-environment variables, and develop deployment-oriented decision-support tools for power infrastructure cost management.

Author contributions

Conceptualization, Cheng Xin and Tianqiong Chen; methodology, Jiaxiang Wen; software, Jiaxiang Wen; validation, Hongda Chen, Zhikai Yang and Huijuan Huo; formal analysis, Jiyuan Zhang; investigation, Jing Duan; resources, Cheng Xin; data curation, Zihao Sun; writing—original draft preparation, Jiaxiang Wen; writing—review and editing, Tianqiong Chen; visualization, Jiaxiang Wen; supervision, Cheng Xin; project administration, Tianqiong Chen; funding acquisition, Tianqiong Chen. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was funded by the Science and Technology Project of the Headquarters of State Grid Corporation of China, grant number 521999250015-182-ZN.

Use of Generative-AI tools declaration

The authors declare that generative AI tools were used solely to polish sentences, correct grammatical errors, and improve the linguistic clarity of this manuscript. All research content is original, and the authors take full responsibility for the final content of the article.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. K. Sun, D. Zhang, J. Wang, W. Mao, Data-driven and physically informed power grid dispatch decision-making method, *Sustain. Energy Grids Netw.*, **42** (2025), 101644. https://doi.org/10.1016/j.segan.2025.101644
2. O. Bertozzi, G. A. Ooi, M. A. Aftab, N. Mohammed, M. S. Venkatraman, G. P. Adam, et al., Data-driven planning of mixed-generation power systems: Towards 100% RES-based grids, *Int. J. Electr. Power Energy Syst.*, **172** (2025), 111207. https://doi.org/10.1016/j.ijepes.2025.111207
3. Z. Ge, C. Yan, J. Guo, C. Zhang, Evaluating the embodied carbon emissions of power infrastructure construction towards carbon neutrality: A study of Fujian Province, China, *Energy Strategy Rev.*, **64** (2026), 102096. https://doi.org/10.1016/j.esr.2026.102096
4. S. Y. Teng, C. C. V. Nooten, J. M. V. Doorn, A. Ottenbros, M. A. J. Huijbregts, J. J. Jansen, Near real-time predictions of renewable electricity production at substation level via domain adaptation zero-shot learning in sequence, *Renew. Sustain. Energy Rev.*, 186 (2023), 113662. https://doi.org/10.1016/j.rser.2023.113662
5. A. Gao, J. Zheng, F. Mei, Y. Liu, Toward intelligent demand-side energy management via substation-level flexible load disaggregation, *Appl. Energy*, **367** (2024), 123361. https://doi.org/10.1016/j.apenergy.2024.123361

6. M. Alipour, J. Aghaei, M. Norouzi, T. Niknam, S. Hashemi, M. Lehtonen, A novel electrical net-load forecasting model based on deep neural networks and wavelet transform integration, *Energy*, **205** (2020), 118106. https://doi.org/10.1016/j.energy.2020.118106
7. B. A. Atalay, K. Zor, An innovative approach for forecasting hydroelectricity generation by benchmarking tree-based machine learning models, *Appl. Sci.*, 15 (2025), 10514. https://doi.org/10.3390/app151910514
8. İ. Levent, G. Şahin, G. Işık, W. G. J. H. M. V. Sark, Comparative analysis of advanced machine learning regression models with advanced artificial intelligence techniques to predict rooftop PV solar power plant efficiency using indoor solar panel parameters, *Appl. Sci.*, **15** (2025), 3320. https://doi.org/10.3390/app15063320
9. Y. Jiang, J. Hu, S. Mei, Y. Zhou, Power engineering cost prediction model based on BP neural network, *Procedia Comput. Sci.*, **261** (2025), 1057–1064. https://doi.org/10.1016/j.procs.2025.04.685
10. H. Lu, M. Fang, Y. Zhang, X. Li, BayesShrink threshold estimation-based multi-metric space clustering algorithm for power grid project cost data, *Results Eng.*, **29** (2026), 109134. https://doi.org/10.1016/j.rineng.2026.109134
11. C. Wang, J. Qiao, Construction project cost prediction method based on improved BiLSTM, *Appl. Sci.*, **14** (2024), 978. https://doi.org/10.3390/app14030978
12. A. Mahmoodzadeh, H. R. Nejati, M. Mohammadi, H. Hashim Ibrahim, M. Khishe, S. Rashidi, et al., Developing six hybrid machine learning models based on Gaussian process regression and meta-heuristic optimization algorithms for prediction of duration and cost of road tunnels construction, *Tunn. Undergr. Space Technol.*, **130** (2022), 104759. https://doi.org/10.1016/j.tust.2022.104759
13. L. Chen, C. Xu, W. H. Lim, A. Sharma, S. S. Tiang, K. S. Chong, et al., Transparent and reliable construction cost prediction using advanced machine learning and explainable AI, *Eng. Sci. Technol. Int. J.*, **70** (2025), 102159. https://doi.org/10.1016/j.jestch.2025.102159
14. F. Antoniou, K. Konstantinidis, Client-oriented highway construction cost estimation models using machine learning, *Appl. Sci.*, **15** (2025), 10237. https://doi.org/10.3390/app151810237
15. U. Park, Y. Kang, H. Lee, S. Yun, A stacking heterogeneous ensemble learning method for the prediction of building construction project costs, *Appl. Sci.*, **12** (2022), 9729. https://doi.org/10.3390/app12199729
16. Y. S. Ko, B. G. Lee, A study on machine learning-based cost estimation models for AI training data construction, *Appl. Sci.*, **16** (2026), 2891. https://doi.org/10.3390/app16062891
17. H. Borchani, G. Varando, C. Bielza, P. Larrañaga, A survey on multi-output regression, *WIREs Data Min. Knowl. Discov.*, **5** (2015), 216–233. https://doi.org/10.1002/widm.1157
18. S. L. Wickramasuriya, G. Athanasopoulos, R. J. Hyndman, Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization, *J. Am. Stat. Assoc.*, **114** (2019), 804–819. https://doi.org/10.1080/01621459.2018.1448825
19. M. Waqar, Y. W. Kim, Y. C. Byun, A hybrid deep learning framework for multivariate energy forecasting and peak load prediction in electric vehicle charging infrastructure, *Appl. Energy*, **402** (2026), 126964. https://doi.org/10.1016/j.apenergy.2025.126964
20. W. Sfar, L. Amhaimar, A. Khalidi, B. Talbi, A hybrid long-term photovoltaic power prediction model integrating a BiLSTM network with residual correction via CatBoost, *Results Eng.*, **29** (2026), 108898. https://doi.org/10.1016/j.rineng.2025.108898

21. Z. Zhou, Y. Dai, M. Leng, A photovoltaic power forecasting framework based on attention mechanism and parallel prediction architecture, *Appl. Energy*, **391** (2025), 125869. https://doi.org/10.1016/j.apenergy.2025.125869
22. H. Liu, H. Wu, H. Jin, Y. He, Adaptive forecasting of photovoltaic power based on dual-type models' ensemble and online error correction, *Appl. Energy*, **408** (2026), 127397. https://doi.org/10.1016/j.apenergy.2026.127397
23. L. Zhang, D. Jánošík, Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches, *Expert Syst. Appl.*, **241** (2024), 122686. https://doi.org/10.1016/j.eswa.2023.122686
24. M. Shafiuzzaman, M. Safayet Islam, T. M. Rubaith Bashar, M. Munem, M. Nahiduzzaman, M. Ahsan, et al., Enhanced very short-term load forecasting with multi-lag feature engineering and Prophet-XGBoost-CatBoost architecture, *Energy*, **335** (2025), 137981. https://doi.org/10.1016/j.energy.2025.137981
25. Z. Colak, Hybrid machine learning approach in forecasting renewable energy production: Extra Trees, CatBoost and LightGBM-based stacking model, *Renew. Energy*, **260** (2026), 125167. https://doi.org/10.1016/j.renene.2025.125167
26. W. Xiang, P. Xu, J. Fang, Q. Zhao, Z. Gu, Q. Zhang, Multi-dimensional data-based medium- and long-term power-load forecasting using double-layer CatBoost, *Energy Rep.*, **8** (2022), 8511–8522. https://doi.org/10.1016/j.egyr.2022.06.063
27. S. Ruan, Y. Chen, G. Lu, Z. Li, S. Chen, C. Wang, et al., Multifactor interpretability method for offshore wind power output prediction based on TPE-CatBoost-SHAP, *Comput. Electr. Eng.*, **123** (2025), 110081. https://doi.org/10.1016/j.compeleceng.2025.110081
28. X. Guo, P. Zeng, X. Xiong, G. Wang, Y. Cui, Short-term wind power forecasting methods based on machine learning: A review and case study, *Energy Rep.*, **14** (2025), 3753–3782. https://doi.org/10.1016/j.egyr.2025.10.040
29. A. Wang, Q. Yu, J. Wang, X. Yu, Z. Wang, Z. Hu, Electric load forecasting based on deep ensemble learning, *Appl. Sci.*, **13** (2023), 9706. https://doi.org/10.3390/app13179706
30. Y. Zhang, Z. Wang, M. Xi, Y. Zhao, J. Liu, Data-driven prediction of residual flexural capacity in corroded RC beams using PSO and GA-optimized CatBoost ensemble models, *Eng. Res. Express*, 7 (2025), 035129. https://doi.org/10.1088/2631-8695/adfcb3
31. Y. Zhang, Z. Wang, J. Liu, Y. Li, Z. Huang, X. Yu, A metaheuristic-driven categorical boosting framework with interpretability for high-precision prediction of mechanical properties in corroded reinforced concrete beams, *Eng. Appl. Artif. Intell.*, **163** (2026), 112804. https://doi.org/10.1016/j.engappai.2025.112804
32. Q. Xu, C. Zhang, S. Geng, S. Wang, J. Li, J. Liu, et al., An explainable ensemble learning framework for flexible pavement roughness prediction under multi-climate stressors, *Case Stud. Constr. Mater.*, **23** (2025), e05402. https://doi.org/10.1016/j.cscm.2025.e05402
33. S. Xiong, C. Nie, Y. Lu, J. Zheng, Research on price prediction of construction materials based on VMD-SSA-LSTM, *Appl. Sci.*, **15** (2025), 2005. https://doi.org/10.3390/app15042005