Evaluating Predictive Models with Ensemble Learning Methods for Construction Progress and Defect Analysis

Ching-Lung Fan[1]

Highlights

- Assessed the performance of four ensemble learning methods—Random Forest, AdaBoost, Gradient Boosting, and Decision Tree—in predicting construction progress.
- Utilized the Public Construction Intelligence Cloud (PCIC) dataset to establish links between defect patterns and construction progress, providing a robust foundation for machine learning models.
- Identified AdaBoost as the most effective classifier, achieving a precision, recall, and F1-score of 91.2%, showcasing its capability in handling multifaceted construction datasets.
- Emphasized the value of ensemble learning for proactive construction management, enabling better defect detection and progress monitoring to enhance project outcomes.

ABSTRACT

The construction industry has historically faced challenges in predicting project progress and managing defects effectively. Traditional methods, such as statistical models, often produce biased results due to their reliance on predefined assumptions. In contrast, machine learning (ML) models offer significant advantages in handling complex datasets with multiple attributes. ML models can identify critical features that influence construction performance without being constrained by distribution or collinearity effects. This study leverages the extensive Public Construction Intelligence Cloud (PCIC) dataset, enabling ML models to uncover hidden patterns related to deficiencies and construction progress, thereby supporting decision-making in construction management. Supervised learning classifiers, including Decision Tree (DT), Random Forest (RF), AdaBoost, and Gradient Boosting, are employed to analyze data collected from construction sites. The findings indicate that ML models can identify correlations and trends within large datasets that traditional methods might overlook, thereby enhancing predictive capabilities and providing actionable insights for construction management.

Keywords: Machine learning, Ensemble learning method, Construction progress, Defects

1. Introduction

[1] Ching-Lung Fan, Associate Professor, Department of Civil Engineering, Republic of China Military Academy, No. 1, Weiwu Rd., Fengshan, Kaohsiung 83059, Taiwan, e-mail: p93228001@ntu.edu.tw (ORCID: 0000-0001-9022-120X).

Construction project documents, such as construction logs, contain significant information and implicit rules. These documents assist project managers in identifying the causes and effects of accidents or patterns within a project (Xu et al., 2019). Therefore, analyzing the extensive data recorded on-site for large-scale civil infrastructure projects and extracting critical features is essential for detecting damages. However, data availability and accessibility in the construction industry remain limited. Many records are still paper-based, and database systems are inconsistently utilized, leading to a lack of integrated and structured data sources (Delgado and Oyedele, 2021). Additionally, while the construction industry generates vast amounts of data, accessing this information is challenging due to issues such as inconsistent data structure, missing data, noise, and the high costs of data collection and preprocessing (Yan et al., 2020). These challenges are further exacerbated by the complexity of collection environments, limitations in data collection equipment, and human factors, all of which hinder the accuracy of quality data. Moreover, the scarcity of public data on construction quality makes it difficult to establish comprehensive data sources. Contractors and project owners may also hesitate or refuse to share quality records due to concerns about exposing sensitive information critical to their projects (Luo et al., 2022).

To address these limitations, this study utilized the construction inspection dataset from the Public Construction Intelligence Cloud (PCIC). Established by the Taiwanese government in 1933, the PCIC is designed to ensure the quality of public construction projects through standardized inspections. Experts and scholars perform on-site quality checks using standardized forms, and the results are digitized and stored within the PCIC. This ensures the accessibility and consistency of inspection data across various construction stages. Given the wealth of data stored in the PCIC, it is essential to analyze the relationship between defects and construction progress in-depth. Such analyses can guide the development of effective construction management strategies, ultimately improving construction quality and project outcomes.

Construction project data often involve numerous variables and nonlinear characteristics, posing challenges for traditional statistical methods. These methods are prone to biased results when variables are highly correlated. In contrast, machine learning (ML) models excel at analyzing datasets with multiple attributes and can identify critical features without being affected by data distribution or collinearity (Uddin et al., 2022). ML enables a deeper understanding of complex construction data, allowing for more effective decision-making. By studying patterns within the data, ML models can build predictive frameworks or extract valuable insights, making them adaptable to various datasets (Ghoddusi et al., 2019). Leveraging ML for construction site data analysis offers opportunities to optimize performance across planning, design, safety, quality, scheduling, and cost management (Ajayi et al., 2019). With the availability of large-scale historical data, ML addresses scalability issues and enhances progress tracking and decision-making capabilities (Amer and Golparvar-Fard, 2021).

To ensure timely and high-quality project completion, it is crucial to adopt advanced technologies for predicting construction progress. A thorough understanding of how deficiencies impact construction stages is key to achieving this goal. ML models excel at uncovering trends and correlations within large datasets that traditional methods might overlook, thereby enhancing predictive accuracy. These insights enable the development of actionable strategies for construction management, creating new value applications. By utilizing the extensive PCIC dataset, this study demonstrates how ML models can effectively learn hidden patterns in deficiencies and construction progress, ultimately supporting informed decision-making in construction management.

2. Research Method

The supervised learning classifiers employed in this study include Decision Tree (DT), Random Forest (RF), AdaBoost, and Gradient Boosting. This section outlines the underlying principles of these classifiers and details the data preprocessing and processing steps specific to each classifier.

2.1. Decision tree

A Decision Tree (DT) is a tree-like structure introduced by Breiman et al. (1984), consisting of nodes, branches, and leaf nodes. Each internal node represents a feature (or attribute), each branch corresponds to a decision rule, and each leaf node denotes an outcome or class label. At each node, the DT selects the optimal splitting criterion—typically a threshold for a feature—to maximize the purity of the resulting subsets with respect to the target variable. The goal is to group samples in a way that ensures each subset predominantly belongs to a single class, thereby enhancing differentiation.

DTs commonly use measures such as entropy or Gini impurity to quantify impurity. For instance, entropy is zero when all instances in a subset belong to a single class, indicating maximum purity. A key advantage of DTs is their ability to capture non-linear relationships in data without relying on assumptions about its distribution. However, they are prone to overfitting, especially when the tree grows too deep, leading to overly complex models. Techniques such as pruning, which removes branches that contribute minimally to predictive performance, can effectively reduce overfitting. While DTs are simple and highly interpretable, their performance can be unstable, as small variations in the data may result in entirely different tree structures. This sensitivity underscores the importance of using DTs judiciously, particularly in scenarios where data variability is high.

2.2. Random forest

Random Forest (RF) is an ensemble learning method designed to address the limitations of individual decision trees by aggregating predictions from multiple trees to enhance accuracy and stability (Breiman, 2001). Each tree in an RF is constructed using a bootstrap sample of the original dataset, and at each node, splits are determined based on the best feature from a randomly selected subset rather than the overall best feature. This incorporation of randomness increases diversity among trees, thereby improving the model's robustness and generalization ability.

RF combines individual tree predictions using majority voting for classification tasks or averaging for regression tasks. Since the construction of each tree is independent, RF is inherently robust against overfitting, particularly in large datasets. The method is notable for its adaptability, computational efficiency, and simplicity. Moreover, RF provides an internal estimate of its generalization error through out-of-bag (OOB) error estimation, which leverages the data samples excluded from the bootstrap process for validation. This feature makes RF a reliable and efficient choice for a wide range of machine learning applications.

2.3. AdaBoost

The weight assigned to each weak classifier is determined by its error rate, with higher weights given to classifiers that achieve lower errors. Through this mechanism, AdaBoost effectively handles complex datasets while reducing the risk of overfitting. Its adaptability allows it to perform well across various data distributions, making it a robust choice for diverse applications. The algorithm iteratively generates a series of sub-models. Initially, the first sub-model is trained on the original dataset. Afterward, the dataset is reweighted based on the predictions of this sub-model, increasing the weights of misclassified samples and decreasing those of correctly classified ones. This process continues for a predefined number of iterations, creating distinct sub-models due to the varying sample weights during training. The final AdaBoost model aggregates the predictions of all sub-models using weighted voting, producing a comprehensive and powerful classification or regression outcome.

2.4. Gradient Boosting

The basic idea of Gradient Boosting is to sequentially generate multiple weak learners, each tasked with fitting the negative gradient of the loss function from the previous accumulated model. This process iteratively reduces the cumulative model loss in the direction of the negative gradient. In neural network training, gradient descent optimizes parameters by calculating the gradient of the loss function with respect to model parameters. Similarly, in Gradient Boosting, each weak learner fits the gradient of the loss function with respect to the accumulated model. The weak learner is then added to the accumulated model, gradually reducing its loss.

Unlike AdaBoost, which adjusts sample weights to focus on misclassified data, Gradient Boosting minimizes the loss function directly through gradient descent. Each new model fits the residual error of the previous model, progressively improving predictions. The key strength of Gradient Boosting lies in its ability to focus on data points poorly handled by previous iterations. This iterative refinement enables Gradient Boosting to capture complex patterns and achieve high accuracy. While computationally intensive, it remains widely adopted for its superior predictive performance across classification and regression tasks.

2.5. Data preprocessing and research process

The PCIC dataset used in this study comprises data from 1,015 construction projects, with 499 recorded defects distributed across four categories: Management (113 defects), Quality (356 defects), Schedule (10 defects), and Design (20 defects). Additionally, the dataset includes 6,615 defect frequencies and two categories of construction progress status: "ahead of schedule" and "behind schedule." This comprehensive dataset undergoes preprocessing to ensure it is suitable for training machine learning (ML) classifiers to predict construction progress effectively.

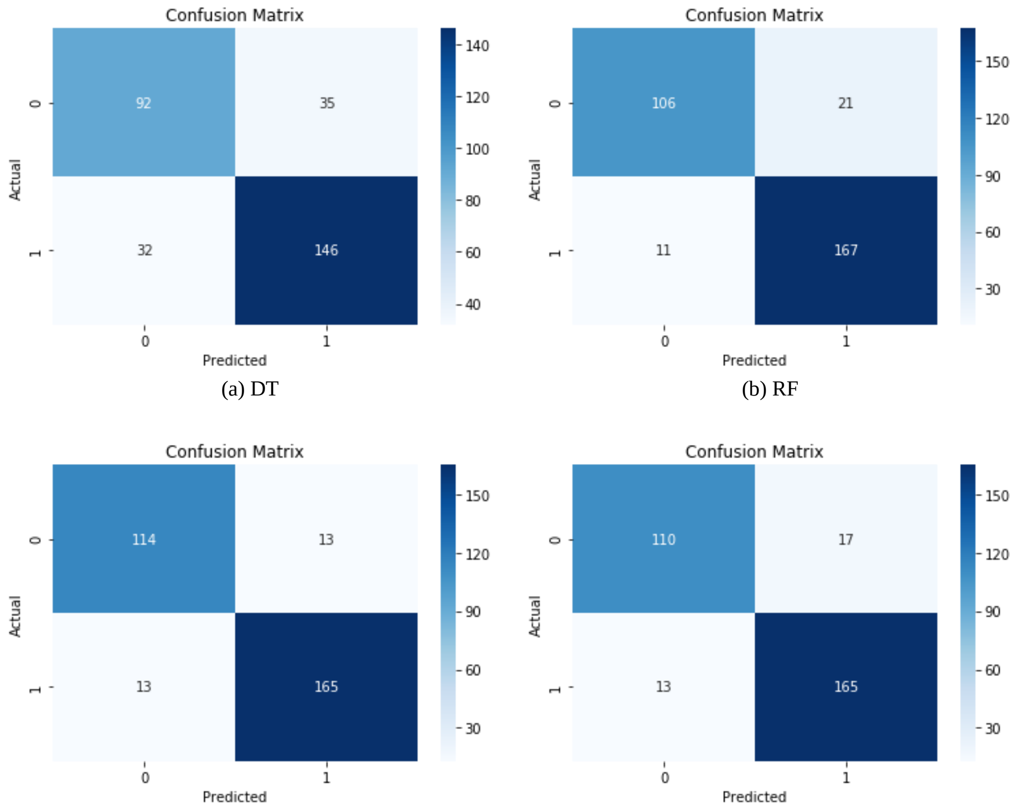
The success of ML applications is heavily dependent on the quality of the input data. As highlighted by Na et al. (2023), the value derived from ML models is significantly influenced by the completeness, accuracy, and reliability of the data. High-quality training data ensures that learning objectives are achieved effectively (Alanne, 2021). For this study, the PCIC dataset provides a robust foundation for supervised learning, linking defect items (independent variables) with construction progress status (dependent variables).

To predict construction progress, four ML classifiers are employed and evaluated using metrics such as confusion matrices, which provide insights into the models' accuracy and reliability. Supervised learning requires a well-structured dataset where independent variables are mapped to corresponding labels. Once training is complete, the models establish a predictive mapping, enabling them to infer progress status for unseen combinations of defect data (Alawadi et al., 2020).

This study categorizes the progress of construction projects into two types based on a comparison of scheduled and actual progress. By leveraging ML classifiers to analyze defect data, the models offer actionable insights that support proactive decision-making. This capability is essential for addressing potential issues early, improving project management efficiency, and ensuring successful construction outcomes.

3. Results

This study employs multiple ML classifiers, including Decision Trees (DT), Random Forest (RF), AdaBoost, and Gradient Boosting, to predict construction progress. The performance of these classifiers is evaluated using confusion matrix analysis, which provides a detailed breakdown of correct and incorrect predictions, offering insights into the model's strengths and weaknesses in different scenarios. The confusion matrix helps understand the distribution of true positives, false positives, true negatives, and false negatives, and can identify areas where the model may need improvement. For example, if the model frequently misclassifies certain progress stages or defect types, targeted data augmentation or model tuning could address these issues. Accuracy, a key metric for evaluating model performance, reflects the proportion of correct predictions. It is defined as the number of correctly predicted samples divided by the total number of samples, expressed as a percentage. Higher accuracy indicates better model performance in correctly identifying progress stages and defects.



(a) DT

(b) RF

(c) AdaBoost

(d) Gradient Boosting

Fig. 1. Confusion matrices of four ML classifiers for construction progress.

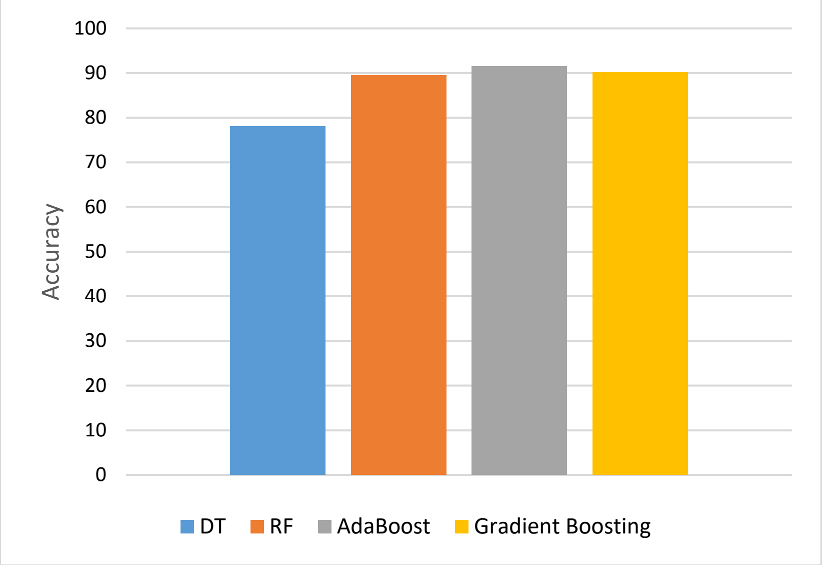


Fig. 2. Accuracy of four ML classifiers for construction progress.

Among the classifiers, AdaBoost stands out with its impressive classification performance, correctly predicting 279 instances of construction progress, as shown in Fig. 1. This demonstrates AdaBoost's ability to handle complex data and make precise predictions, making it a valuable tool for construction project management. In conclusion, the use of these ML classifiers, particularly ensemble learning algorithms, shows significant promise in enhancing the accuracy and reliability of predictive models for construction project management. The findings highlight the potential of ML to provide actionable insights and improve decision-making in the construction industry. Fig. 2 presents the accuracy of the four ML classifiers used in this study. The results indicate that ensemble learning algorithms (RF, AdaBoost, and Gradient Boosting) perform exceptionally well, achieving an average accuracy of 90.38% for construction progress. This high level of accuracy underscores the robustness of ensemble methods in aggregating the predictive power of multiple models to improve overall performance.

The most important consideration in model development is to evaluate its performance in an unbiased manner. To achieve this, the dataset is split into two parts: a training set, which comprises 70% of the data, and a testing set, which comprises the remaining 30%. The training set is used for model development, allowing the algorithm to learn from the data, while the testing set is reserved for performance evaluation, providing an unbiased assessment of how well the model generalizes to unseen data. Training and testing are essential processes in implementing supervised ML techniques. They ensure that the model's predictive capabilities are rigorously tested and validated. During the training phase, the model learns the underlying patterns and relationships within the data, while in the testing phase, its performance is evaluated using various metrics to determine accuracy and reliability.

In this study, AdaBoost achieves the best prediction results in terms of construction progress, demonstrating superior performance across multiple evaluation metrics. Specifically, AdaBoost attains a Precision of 91.2%, a Recall of 91.2%, and an F1-score of 0.912, as detailed in Table 1. Precision measures the proportion of true positive predictions among all positive predictions, Recall measures the proportion of true positives identified correctly out of all actual positives, and the F1-score provides the harmonic mean of

Precision and Recall, offering a balanced measure of the model's performance. Table 1 summarizes the performance evaluation of the four ML classifiers for predicting construction progress. Each classifier's performance is quantified using Precision, Recall, and F1-score. DT achieved a Precision of 77.4%, a Recall of 77.2%, and an F1-score of 0.773. RF demonstrated stronger performance with a Precision of 89.7%, a Recall of 88.6%, and an F1-score of 0.891. Gradient Boosting also showed high performance, with a Precision of 90.0%, a Recall of 89.7%, and an F1-score of 0.898.

The superior performance of AdaBoost highlights its effectiveness in handling the complexities of construction progress data. Its ability to combine weak classifiers to form a strong classifier results in enhanced predictive accuracy. The evaluation metrics underscore the robustness of AdaBoost, making it a reliable tool for construction project management. Furthermore, these performance metrics are crucial for identifying the strengths and weaknesses of each model. For instance, while DTs may be simpler and faster, their lower Precision and Recall compared to ensemble methods like RF and AdaBoost suggest they may not be as effective for this specific application. In contrast, the higher scores of ensemble methods indicate their ability to capture more intricate patterns within the data, leading to better generalization and prediction accuracy.

Through the rigorous analysis and tuning of ML algorithms, this study highlights the transformative potential of ML in addressing complex construction prediction challenges. It emphasizes how advanced ML techniques can be leveraged to enhance project management practices, offering a path toward more efficient and successful construction project execution.

Table 1. Performance evaluation of the four ML classifiers for construction progress.

Machine learning	Precision	Recall	F1-score
DT	0.774	0.772	0.773
RF	0.897	0.886	0.891
AdaBoost	0.912	0.912	0.912
Gradient Boosting	0.90	0.897	0.898

4. Conclusion

ML is fundamentally a data-driven field that relies heavily on large volumes of training data to achieve optimal performance for practical deployment. The essence of ML lies in its ability to construct predictive models based on experience and patterns observed in data. These models are built by computers using various algorithms and are designed to perform predictive analytics across a wide range of domains. In the field of project management, ML techniques are frequently employed to analyze project data, identify patterns, and generate predictions that support decision-making processes. Specifically, ML integrates concepts from computer science, statistics, and data science to tackle complex prediction tasks. This integration allows ML to provide substantial insights and support evidence-based decision-making by uncovering patterns and relationships within data that may not be apparent through traditional methods.

In this study, the construction inspection dataset from the PCIC is utilized to enhance ML model performance. This process involves hyperparameter tuning and training across four different ML algorithms to develop optimal prediction models. The primary goal is to construct models that can accurately forecast construction progress based on the dataset, which includes a diverse array of defect items and associated frequencies. The results of the study reveal that the AdaBoost algorithm exhibits superior performance in predicting construction progress, achieving an impressive accuracy rate of 91.2%. AdaBoost, an ensemble learning method that combines multiple weak classifiers to form a robust predictive model, demonstrates its effectiveness in handling the complexities of construction data.

The insights gained from ML models are instrumental for project managers, as they enable them to better identify significant defects and potential issues through detailed analysis of prediction errors. By understanding the nature of these errors, managers can implement more effective construction management strategies, ensuring that projects adhere to schedules and meet quality standards. The application of ML in construction project management not only enhances the accuracy of progress predictions but also supports more informed decision-making, ultimately leading to improved project outcomes.

5. Key References

Xu, Y., Wei, S. Y., Bao, Y. Q., & Li, H. (2019). Automatic seismic damage identification of reinforced concrete columns from images by a region-based deep convolutional neural network. *Structural Control and Health Monitoring, 26 (3)*, e2313.

Delgado, J. M. D., & Oyedele, L. (2021). Deep learning with small datasets: using autoencoders to address limited datasets in construction management. *Applied Soft Computing, 112*, 107836.

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction, 119*, 103331.

Luo, H., Lin, L., Chen, K., Antwi-Afari, M. F., & Chen, L. (2022). Digital technology for quality management in construction: A review and future research directions. *Developments in the Built Environment, 12*, 100087.