infrastructures

IMPACT FACTOR 3.6

CITESCORE 5.7

Article

Application of Gradient Boosting for Predicting Cost Overrun in Construction Projects

Johanis Anggawan, Agustinus Purna Irawan and Mega Waty

Topic

Construction Project Management and Infrastructure Sustainability

Edited by

Prof. Dr. Tarek Zayed, Dr. Mohamed Hussein, Dr. Sherif Abdelkhalek, Dr. Eslam Saleh and Dr. Abdul Hannan Qureshi

MDPI

https://doi.org/10.3390/infrastructures11090328

infrastructures

MDPI

Article

Application of Gradient Boosting for Predicting Cost Overrun in Construction Projects

Johanis Anggawan [1,*], Agustinus Purna Irawan [2] and Mega Waty [1]

[1] Fakultas Teknik Program Studi Doktor Ilmu Teknik Sipil, Universitas Tarumanagara, Kota Jakarta Barat 11440, DKI Jakarta, Indonesia; mega@ft.untar.ac.id
[2] School of STEM, Universitas Prasetiya Mulya, Jakarta 12430, DKI Jakarta, Indonesia; agustinus@prasetiyamulya.ac.id
* Correspondence: johanis.328211004@stu.untar.ac.id

Abstract

Cost overrun remains a major challenge in construction project management, particularly in regions characterized by geographical constraints, logistical complexity, and high project uncertainty. Conventional statistical approaches often struggle to capture nonlinear relationships among factors associated with construction cost overruns. This study proposes a hybrid framework integrating Structural Equation Modeling–Partial Least Squares (SEM-PLS) and Gradient Boosting to predict a latent Cost Overrun construct in construction projects in Southwest Papua Province, Indonesia. SEM-PLS was employed to assess the validity and reliability of the research constructs, comprising 24 predictor indicators across three explanatory constructs and three outcome indicators for the latent Cost Overrun construct. The validated indicator-level latent scores were used as input features for the Gradient Boosting model, while the latent Cost Overrun construct was used as the prediction target. The model achieved a mean absolute error (MAE) of 1.7838, a root mean square error (RMSE) of 2.2079, a mean absolute percentage error (MAPE) of 19.10%, and an R^2 of 0.6173. Feature importance analysis indicated that Risk Factors contributed most strongly (53.64%), followed by Social Factors (26.03%) and Geographical Factors (20.33%). The proposed framework provides a structured approach for risk-oriented construction cost management in high-risk regions.

Keywords: cost overrun; construction projects; Gradient Boosting; SEM-PLS; Southwest Papua

Check for updates

Academic Editors: Tarek Zayed, Mohamed Hussein, Sherif Abdelkhalek, Eslam Saleh and Abdul Hannan Qureshi

Received: 14 July 2026
Revised: 14 August 2026
Accepted: 19 August 2026
Published: 15 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Cost overrun is one of the major challenges in construction project management and represents a critical indicator of project success. Cost overrun occurs when the actual project expenditure exceeds the estimated cost established during the initial planning stage or the agreed contract value [1]. This phenomenon not only reduces contractor profitability but also affects investment efficiency, project financing sustainability, and the effectiveness of public expenditure in infrastructure development [2]. Previous studies have demonstrated that cost overruns are a global and persistent issue across various types of construction projects, including roads, bridges, buildings, ports, airports, energy infrastructure, and public transportation systems [3,4]. More than 90% of public infrastructure projects analyzed in previous studies experienced cost overruns with substantial deviations from their original estimates. Similar conditions have also been observed in developing countries,

where additional challenges such as material price volatility, supply chain limitations, and logistical and geographical constraints further complicate project delivery [5]. In large-scale infrastructure projects, cost overruns are frequently triggered by project uncertainties, scope changes, fluctuations in material prices, and the limited accuracy of cost estimation during the early planning stage [6].

In developing countries, the complexity of cost overruns becomes even greater due to the interaction of economic, technical, managerial, and external factors that dynamically influence project performance. Material price instability, variations in labor productivity, procurement delays, design changes during project execution, unforeseen site conditions, and regulatory changes have consistently been identified as major contributors to construction cost overruns [7]. Furthermore, geographical conditions and limited logistics infrastructure in certain regions may increase cost uncertainty and project implementation risks.

Within the Indonesian context, these challenges become more pronounced in projects located in regions characterized by limited accessibility and high dependence on material transportation from other areas. Under such conditions, the relationships among cost overrun factors become increasingly nonlinear and difficult to explain using conventional cost estimation approaches, which typically assume linear relationships among variables. In practice, construction project data are often influenced by multidimensional interactions among technical, economic, managerial, and environmental factors that evolve dynamically throughout the project lifecycle [8–10].

The complexity of these interactions makes cost estimation and prediction increasingly difficult when using conventional statistical approaches based on linear assumptions. In reality, the relationships among technical, economic, managerial, and environmental variables are often nonlinear and dynamic, resulting in patterns that cannot be adequately captured by traditional statistical methods [11]. Consequently, linear regression-based models and conventional statistical approaches frequently exhibit limitations in modeling complex relationships and producing adaptive cost predictions under changing project conditions [12]. Although these methods may perform adequately for datasets with limited variables and relatively simple relationships, their predictive capability tends to decline when applied to construction projects characterized by heterogeneous datasets and complex variable interactions [13].

Recent advances in computational technology and the increasing availability of large-scale construction project data have accelerated the adoption of machine learning techniques for cost prediction and decision support in construction management. Various algorithms, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), have been employed to improve cost prediction accuracy and have demonstrated superior performance compared with conventional statistical approaches [14–18]. In particular, ensemble learning methods have shown a stronger capability to capture nonlinear relationships, reduce the risk of overfitting, and identify hidden patterns within complex construction datasets [19]. Recent studies in infrastructure engineering have further demonstrated that boosting-based algorithms can significantly improve predictive performance across various civil engineering and infrastructure management applications [20–22]. Consequently, machine learning approaches based on ensemble learning are increasingly recognized as promising alternatives for achieving more accurate and adaptive cost overrun prediction.

Despite these advances, several research gaps remain to be addressed. First, most previous studies have focused primarily on construction cost estimation rather than cost overrun prediction during project execution [23]. Second, many predictive models have been developed using all available variables without systematic construct validation or feature selection procedures, potentially resulting in unstable models with limited inter-

pretability [24]. Third, studies investigating the application of Gradient Boosting algorithms for cost overrun prediction in construction projects, particularly in developing countries characterized by high geographical, logistical, and project risk complexity, remain relatively limited [25]. Furthermore, previous research has predominantly emphasized predictive accuracy while paying less attention to model interpretability and the identification of dominant factors contributing to project cost overruns [26].

To address these limitations, this study proposes a cost overrun prediction framework that integrates Structural Equation Modeling–Partial Least Squares (SEM-PLS) as a construct validation and feature selection mechanism with Gradient Boosting as the primary prediction model. SEM-PLS is employed to identify factors that significantly influence cost overruns, whereas Gradient Boosting is utilized to model nonlinear relationships among variables and generate more accurate predictions. The proposed framework further evaluates model performance using several prediction metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2), while also comparing the performance of Gradient Boosting with several alternative machine learning algorithms.

The contributions of this study can be summarized in three aspects. First, the study introduces an integrated SEM-PLS and Gradient Boosting framework to establish a more systematic feature screening based on causal relationships among variables. Second, the study extends the application of ensemble learning techniques for cost overrun prediction in developing countries characterized by high uncertainty and logistical complexity. Third, the study provides a practical contribution by developing a predictive model that can serve as a decision support tool for project owners, contractors, and government agencies in construction cost control and infrastructure investment planning.

Based on these considerations, the objectives of this study are fourfold: (1) to identify factors that significantly influence construction cost overruns; (2) to develop a cost overrun prediction model using the Gradient Boosting algorithm; (3) to evaluate model performance using multiple prediction metrics; and (4) to identify dominant factors affecting cost overruns through feature importance analysis. The findings are expected to contribute both theoretically to the advancement of machine learning-based cost prediction models and practically to improving the effectiveness of cost control in construction and infrastructure projects.

2. Materials and Methods

2.1. Research Framework

This study adopts a quantitative data-driven approach to develop a cost overrun prediction model for construction projects using the Gradient Boosting algorithm. The research framework was designed to integrate variable validation, feature selection, predictive model development, and model performance evaluation into a systematic analytical workflow. The first stage of the study involved the collection of historical construction project data encompassing various technical, economic, managerial, and environmental characteristics that potentially influence the occurrence of cost overruns. Subsequently, data cleaning and preprocessing procedures were conducted to ensure the quality and reliability of the dataset used for model development. These procedures included handling missing values, checking data consistency, and performing data transformation where necessary.

In the next stage, Structural Equation Modeling–Partial Least Squares (SEM-PLS) was employed to validate the research constructs and identify variables that significantly influence cost overruns. This process aimed to establish a more systematic feature selection mechanism based on interrelationships among constructs, thereby improving both the stability and interpretability of the predictive model. Variables that satisfied the validation criteria were subsequently used as input features for the Gradient Boosting model. The

Gradient Boosting algorithm was selected because of its capability to model nonlinear relationships, capture complex interactions among variables, and reduce the risk of overfitting through a boosting-based ensemble learning approach [27]. To achieve optimal predictive performance, hyperparameter optimization was conducted by tuning several key parameters that influence the model learning process.

Model performance was subsequently evaluated using several metrics commonly adopted in construction cost prediction studies, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). In addition, feature importance analysis was performed to identify the dominant factors contributing to cost overruns in construction projects. Overall, the proposed research framework aims to develop a cost overrun prediction model that not only achieves high predictive accuracy but also provides meaningful interpretation of the key factors influencing project cost overruns. Consequently, the proposed model has the potential to serve as a decision support tool for project stakeholders in construction project cost management and control. The overall research framework is presented in Figure 1.

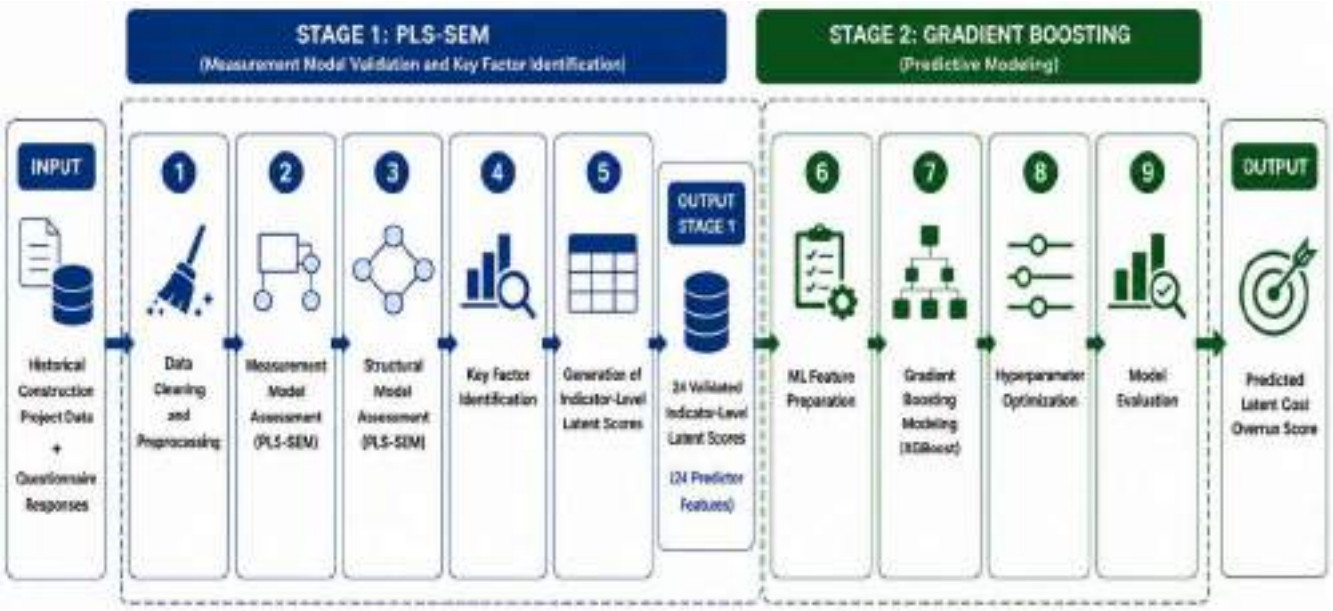


Figure 1. Research framework integrating SEM-PLS and Gradient Boosting.

2.2. Data Collection and Dataset Description

This study employed a quantitative approach based on historical construction project data and survey data to develop a cost overrun prediction model for construction projects in Southwest Papua Province, Indonesia. The research data were collected from construction projects implemented between 2015 and 2025 and were complemented by primary data obtained through questionnaire surveys administered to construction project stakeholders with experience in project execution within the study area. The initial stage of the study involved identifying factors that potentially influence the occurrence of cost overruns based on an extensive literature review and the specific characteristics of the study region. Based on this identification process, the research variables were classified into three main constructs: social factors, geographical factors, and risk factors. In total, the study initially considered 24 indicators representing these three constructs, which served as the basis for constructing the latent variables used in the analysis.

The questionnaire indicators were measured using a Likert scale. Respondents were selected using purposive sampling based on three criteria: having experience in construction projects, being involved in project decision-making or implementation, and understand-

ing construction project cost aspects. The respondents included construction contractors, consultants, and field supervisors involved in construction project activities. A total of 233 questionnaires were distributed, of which 233 were valid and used for analysis, resulting in a 100% response rate.

The social factor construct consisted of four indicators representing the interaction between construction projects and local communities, namely community conflicts, community acceptance, local labor availability, and cultural adaptation. The geographical factor construct comprised four indicators reflecting the physical and accessibility characteristics of the study area, including accessibility, material transportation distance, topographical conditions, and infrastructure limitations. Meanwhile, the risk factor construct included sixteen indicators covering logistics risk, weather risk, regulatory risk, contractual risk, cost control, design changes, labor productivity, construction inflation, and other risks that may arise during project execution. Subsequently, Structural Equation Modeling–Partial Least Squares (SEM-PLS) was employed to assess the measurement and structural models of the research. The SEM-PLS analysis evaluated the validity and reliability of the indicators representing three theoretical constructs, namely Social Factors (X1), Geographical Factors (X2), and Risk Factors (X3). The three constructs were represented by 24 indicators, consisting of four indicators for Social Factors (X1.1–X1.4), four indicators for Geographical Factors (X2.1–X2.4), and sixteen indicators for Risk Factors (X3.1–X3.16). These 24 indicators represent the predictor side of the model, whereas the latent Cost Overrun construct (Y) was measured separately using three outcome indicators (Y.1–Y.3). Thus, the complete SEM-PLS measurement model comprised 27 indicators, while 24 validated predictor indicators were retained as input features for the Gradient Boosting model. The 24 indicators that satisfied the predefined measurement criteria were retained for the subsequent machine learning stage, and their corresponding indicator-level latent scores were used as numerical representations of the validated indicators in the machine learning dataset.

Thus, SEM-PLS served not only to establish the theoretical and empirical validity of the research constructs but also to determine the suitability of the indicators for subsequent predictive modelling. The machine learning stage subsequently used the retained indicator-level latent scores as input features rather than using the three constructs as aggregated input variables. The latent Cost Overrun construct represents a perception-based assessment of the occurrence and severity of construction cost-overrun conditions derived from questionnaire indicators. Accordingly, the Gradient Boosting model predicts the latent Cost Overrun construct score rather than the actual monetary amount or percentage of project cost overrun.

The use of indicator-level latent scores provided numerical representations of the validated indicators for the machine learning model while preserving the information associated with the three theoretical constructs. The predictor features and the prediction target were derived from the same 233 questionnaire observations. Therefore, the Gradient Boosting analysis was treated as a predictive modelling stage based on a predefined hold-out test set rather than as an independent-data validation of the SEM-PLS structural model. Accordingly, the R^2 values obtained from SEM-PLS and Gradient Boosting are not interpreted as evidence that Gradient Boosting is statistically superior to SEM-PLS, but rather as results from two analytical approaches applied to the same survey-derived latent outcome for different analytical purposes. The input features were derived from the indicators that satisfied the SEM-PLS measurement criteria and were organized according to their respective constructs: Social Factors (X1.1–X1.4), Geographical Factors (X2.1–X2.4), and Risk Factors (X3.1–X3.16). The testing set was reserved for final model evaluation and was not used for hyperparameter selection. The training dataset was used to develop the Gradient Boosting model, while the testing dataset was employed to evaluate the model's generalization capability on unseen data. This data partitioning strategy is widely adopted

in machine learning model development to minimize the risk of overfitting and provide a more objective estimation of predictive performance. The main characteristics of the dataset are summarized in Table 1.

Table 1. Characteristics of the Dataset.

Item	Description
Study Area	Southwest Papua Province, Indonesia
Observation Period	2015–2025
Number of Observations	233
Number of Research Constructs	4
Number of Initial Indicators	24
Number of Outcome Indicators	3
Total Measurement Indicators	27
Research Constructs	Social Factor (X1), Geographical Factor (X2), Risk Factor (X3), Latent Cost Overrun construct (Y)
Indicator Structure	X1.1–X1.4, X2.1–X2.4, X3.1–X3.16
Prediction Target	Latent Cost Overrun construct (Y)
Feature Validation	SEM-PLS
Feature Input for ML	Validated indicator-level latent scores
Prediction Algorithm	Gradient Boosting
Training–Testing Ratio	80:20
Training–Testing Split	186 training observations (80%) and 47 testing observations (20%)
Random State	42

2.3. Descriptive Statistics

Descriptive statistics were employed to provide an initial overview of the characteristics of the dataset used for the development of the cost overrun prediction model. This analysis aimed to identify the central tendency, variability, and distribution range of the research variables prior to the machine learning model training process. Since the prediction model was developed using an integrated SEM-PLS and Gradient Boosting approach, descriptive statistics were presented at the latent variable level, comprising Social Factors (X1), Geographical Factors (X2), Risk Factors (X3), and the latent Cost Overrun construct (Y). These construct-level descriptive statistics are presented to summarize the theoretical variables, whereas the machine learning model operates on the validated indicator-level latent scores retained after the SEM-PLS assessment. These variables were represented by latent variable scores generated from the research indicators that satisfied the validity and reliability criteria during the SEM-PLS analysis stage.

Table 2 presents the mean, standard deviation, minimum value, and maximum value for each variable used in the modelling process. The mean value represents the general tendency of the data, whereas the standard deviation indicates the degree of variation among observations. The minimum and maximum values illustrate the distribution range of the dataset used during the model training and evaluation processes.

Table 2. Descriptive Statistics of the Latent Variables.

Variable	Mean	Std. Dev	Min	Max
Social Factor (X1)	3.747	0.892	1.000	5.000
Geographical Factor (X2)	4.185	0.951	1.000	5.000
Risk Factor (X3)	3.983	0.816	1.000	5.000
Latent Cost Overrun construct (Y)	10.416	3.481	3.000	15.000

Based on Table 2, the geographical factor exhibited the highest mean value of 4.185, indicating that most respondents perceived geographical conditions and accessibility limitations as dominant challenges affecting the implementation of construction projects in Southwest Papua. The risk factor recorded a mean value of 3.983 with a relatively low level of variability (SD = 0.816), suggesting a high degree of consistency among respondents regarding the influence of project risks on cost overruns. Meanwhile, the social factor showed a mean value of 3.747, with a slightly higher level of variability compared with the other factors.

The latent Cost Overrun construct had a mean score of 10.416, a standard deviation of 3.481, and a score range between 3 and 15. These results indicate substantial variation in the perceived occurrence and severity of cost-overrun conditions across the observations, providing a suitable basis for developing a machine learning model capable of capturing nonlinear relationships among the latent variables.

As shown in Table 3, most indicator loadings exceeded the minimum recommended threshold of 0.70. Two indicators, X1.3 (0.672) and X3.2 (0.691), were below this threshold but were retained because the corresponding constructs demonstrated satisfactory reliability and convergent validity at the construct level.

Table 3. Outer Loadings of Research Indicators.

Indicator	Social Factor (X1)	Geographical Factor (X2)	Risk Factor (X3)	Latent Cost Overrun (Y)
X1.1	0.843	—	—	—
X1.2	0.749	—	—	—
X1.3	0.672	—	—	—
X1.4	0.715	—	—	—
X2.1	—	0.868	—	—
X2.2	—	0.907	—	—
X2.3	—	0.924	—	—
X2.4	—	0.884	—	—
X3.1	—	—	0.774	—
X3.2	—	—	0.691	—
X3.3	—	—	0.742	—
X3.4	—	—	0.721	—
X3.5	—	—	0.775	—
X3.6	—	—	0.804	—
X3.7	—	—	0.783	—
X3.8	—	—	0.779	—
X3.9	—	—	0.742	—
X3.10	—	—	0.791	—
X3.11	—	—	0.781	—
X3.12	—	—	0.770	—
X3.13	—	—	0.732	—
X3.14	—	—	0.767	—
X3.15	—	—	0.796	—
X3.16	—	—	0.748	—
Y.1	—	—	—	0.888
Y.2	—	—	—	0.863
Y.3	—	—	—	0.838

The cross-loading results in Table 4 show that each indicator has its highest loading on the construct to which it theoretically belongs compared with the other constructs. This pattern indicates that the indicators demonstrate adequate discriminant validity at the indicator level.

Table 4. Cross-Loading of Research Indicators.

Indicator	X1	X2	X3	Y
X1.1	0.843	0.654	0.682	0.595
X1.2	0.749	0.547	0.507	0.382
X1.3	0.672	0.369	0.408	0.287
X1.4	0.715	0.447	0.516	0.325
X2.1	0.605	0.868	0.616	0.492
X2.2	0.600	0.907	0.649	0.430
X2.3	0.665	0.924	0.663	0.489
X2.4	0.636	0.884	0.729	0.447
X3.1	0.607	0.684	0.774	0.352
X3.2	0.536	0.503	0.691	0.311
X3.3	0.563	0.591	0.742	0.415
X3.4	0.545	0.544	0.721	0.413
X3.5	0.541	0.520	0.775	0.518
X3.6	0.563	0.540	0.804	0.494
X3.7	0.569	0.545	0.783	0.438
X3.8	0.591	0.549	0.779	0.428
X3.9	0.546	0.625	0.742	0.418
X3.10	0.507	0.591	0.791	0.447
X3.11	0.518	0.614	0.781	0.378
X3.12	0.539	0.572	0.770	0.433
X3.13	0.524	0.515	0.732	0.445
X3.14	0.555	0.581	0.767	0.448
X3.15	0.598	0.584	0.796	0.449
X3.16	0.613	0.522	0.748	0.432
Y.1	0.451	0.440	0.458	0.888
Y.2	0.457	0.396	0.449	0.863
Y.3	0.546	0.500	0.542	0.838

Discriminant validity was further assessed using the Heterotrait–Monotrait ratio (HTMT). The HTMT results are presented in Table 5.

Table 5. Heterotrait–Monotrait Ratio (HTMT) for Discriminant Validity.

Construct Relationship	HTMT
Social Factor (X1)—Geographical Factor (X2)	0.805
Social Factor (X1)—Risk Factor (X3)	0.833
Geographical Factor (X2)—Risk Factor (X3)	0.797
Social Factor (X1)—Latent Cost Overrun construct (Y)	0.663
Geographical Factor (X2)—Latent Cost Overrun construct (Y)	0.589
Risk Factor (X3)—Latent Cost Overrun construct (Y)	0.622

The HTMT values were 0.805 for the relationship between Social Factor (X1) and Geographical Factor (X2), 0.833 for Social Factor (X1) and Risk Factor (X3), 0.797 for Geographical Factor (X2) and Risk Factor (X3), 0.663 for Social Factor (X1) and the Latent Cost Overrun construct (Y), 0.589 for Geographical Factor (X2) and the Latent Cost Overrun construct (Y), and 0.622 for Risk Factor (X3) and the Latent Cost Overrun construct (Y). All HTMT values were below the 0.90 threshold, indicating adequate discriminant validity among the constructs.

As shown in Table 6, all outer and inner VIF values were below 5, indicating that the measurement and structural models did not exhibit critical multicollinearity.

Table 6. Variance Inflation Factor (VIF) Assessment.

Type	Indicator/Relationship	VIF
Outer VIF	X1.1	1.499
Outer VIF	X1.2	1.462
Outer VIF	X1.3	1.652
Outer VIF	X1.4	1.721
Outer VIF	X2.1	2.440
Outer VIF	X2.2	3.366
Outer VIF	X2.3	3.862
Outer VIF	X2.4	3.035
Outer VIF	X3.1	2.533
Outer VIF	X3.2	2.112
Outer VIF	X3.3	2.519
Outer VIF	X3.4	2.403
Outer VIF	X3.5	2.764
Outer VIF	X3.6	3.452
Outer VIF	X3.7	3.172
Outer VIF	X3.8	2.904
Outer VIF	X3.9	2.174
Outer VIF	X3.10	2.708
Outer VIF	X3.11	2.593
Outer VIF	X3.12	2.573
Outer VIF	X3.13	2.218
Outer VIF	X3.14	3.030
Outer VIF	X3.15	3.403
Outer VIF	X3.16	2.250
Outer VIF	Y.1	2.549
Outer VIF	Y.2	2.335
Outer VIF	Y.3	1.557
Inner VIF	X1 → Y	2.429
Inner VIF	X2 → Y	2.522
Inner VIF	X3 → Y	2.741

2.4. SEM-PLS Validation and Feature Screening

Prior to the development of the Gradient Boosting model, Structural Equation Modeling–Partial Least Squares (SEM-PLS) was employed to assess the validity and reliability of the research constructs and their indicators. The analysis consisted of measurement-model evaluation and structural-model assessment. The measurement model was evaluated using convergent validity, discriminant validity, and reliability criteria to determine whether the indicators adequately represented their respective theoretical constructs. Convergent validity was assessed using outer loadings and Average Variance Extracted (AVE), while discriminant validity was examined using cross-loadings and the Heterotrait–Monotrait ratio (HTMT). Multicollinearity was assessed using both outer and inner Variance Inflation Factor (VIF) values.

The structural model was subsequently assessed to examine the relationships between Social Factors (X1), Geographical Factors (X2), Risk Factors (X3), and the latent Cost Overrun construct (Y). The SEM-PLS assessment involved 24 initial indicators distributed across three theoretical constructs: Social Factors, represented by X1.1–X1.4; Geographical Factors, represented by X2.1–X2.4; and Risk Factors, represented by X3.1–X3.16. The results of the measurement-model assessment were used to determine whether the 24 indicators were sufficiently valid and reliable for subsequent machine learning modelling. All 24 indicators were retained for the machine learning stage based on the overall measurement-model assessment.

For the predictive modelling stage, the validated indicators were represented by their corresponding latent scores generated through SEM-PLS. These indicator-level latent scores were used as machine learning features, while the latent Cost Overrun score was used as the prediction target. Therefore, SEM-PLS did not reduce the 24 indicators into three machine learning input variables. Instead, the three constructs served as the theoretical organization of the indicators, while the validated indicator-level scores provided the features used by the Gradient Boosting model. This integration enables SEM-PLS to provide theoretical and empirical validation before predictive modelling, while Gradient Boosting is subsequently used to capture nonlinear relationships and complex interactions among the validated indicator-level features. The SEM-PLS analysis was conducted using SmartPLS 4.

2.5. Gradient Boosting Model Development

The cost overrun prediction model proposed in this study was developed using the Gradient Boosting algorithm, one of the most widely adopted decision tree-based ensemble learning methods that constructs predictive models sequentially to minimize prediction errors generated in previous iterations [28]. Unlike bagging approaches, which develop multiple models independently and combine their outputs, Gradient Boosting generates a sequence of weak learners in which each newly constructed model focuses on reducing the residual errors produced by the preceding model [29,30].

Mathematically, the Gradient Boosting learning process begins by constructing an initial model based on the average value of the target variable. In each subsequent iteration, the model calculates the residuals or prediction errors generated by the previous iteration and then builds a new decision tree to learn the patterns contained within these residuals. The final prediction is obtained by aggregating the contributions of all decision trees constructed throughout the iterative learning process, as expressed in Equation (1).

$$f_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (1)$$

where:

- $F_m(x)$ represents the prediction model at iteration m;
- $F_{m-1}(x)$ denotes the model generated in the previous iteration;
- $h_m(x)$ represents the weak learner developed to model the residual errors;
- η is the learning rate parameter controlling the contribution of each decision tree to the final prediction model.

The Gradient Boosting algorithm was selected in this study because of its capability to model nonlinear relationships, capture complex interactions among variables, and reduce the risk of overfitting through a sequential boosting mechanism [29–31]. The model was implemented using the GradientBoostingRegressor algorithm from the scikit-learn library. Furthermore, the algorithm provides information regarding the relative importance of each variable through feature importance analysis, thereby enabling the identification of dominant factors contributing to cost overruns in construction projects. The input features used in the prediction model consisted of the validated indicator-level latent scores obtained from the SEM-PLS analysis. These features represented the indicators retained from the three research constructs, namely Social Factors (X1.1–X1.4), Geographical Factors (X2.1–X2.4), and Risk Factors (X3.1–X3.16). The latent Cost Overrun construct (Y) was used as the prediction target. This feature structure allows the Gradient Boosting model to evaluate the predictive contribution of individual indicators while preserving their theoretical association with the three research constructs. The model therefore estimates the latent Cost Overrun construct score from the 24 validated indicator-level latent input features. The final dataset consisted of 233 observations and was divided using hold-out

validation into 80% training data (186 observations) and 20% testing data (47 observations), with a fixed random state of 42. The training set was used for model development and hyperparameter optimization, whereas the testing set was reserved exclusively for the final evaluation of model performance on unseen observations. The architecture of the proposed Gradient Boosting model is presented in Figure 2.

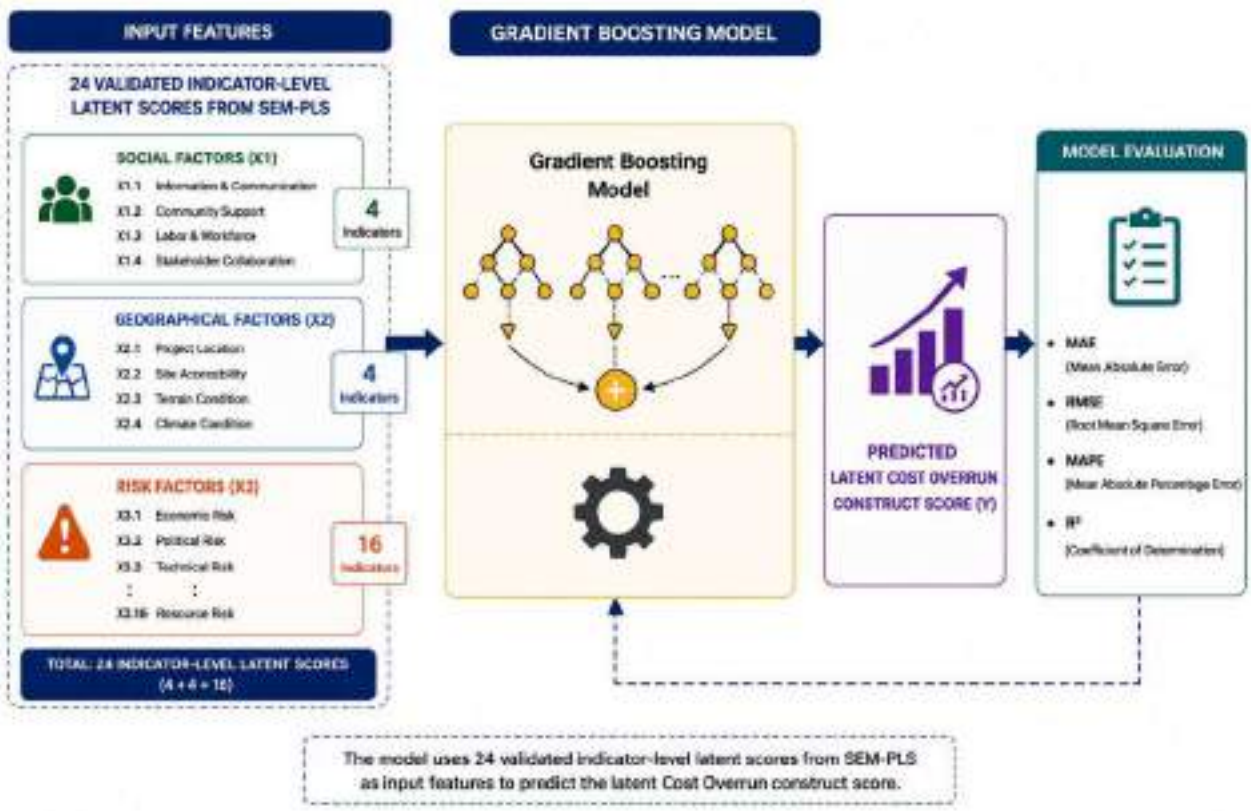


Figure 2. Architecture of the proposed Gradient Boosting model.

2.6. Hyperparameter Optimization

To improve predictive performance and reduce the risk of overfitting, a hyperparameter optimization process was conducted for the Gradient Boosting model. The objective of this process was to identify the combination of parameters that provides the best balance between prediction accuracy and model generalization capability on unseen data.

Several key hyperparameters were optimized in this study, including the number of decision trees (n_estimators), learning rate (learning_rate), maximum tree depth (max_depth), proportion of training samples used in each iteration (subsample), minimum number of samples required to split an internal node (min_samples_split), minimum number of samples required at a leaf node (min_samples_leaf), and the maximum number of features considered at each split (max_features). These hyperparameters were selected because they have a significant influence on model complexity and the predictive performance of the Gradient Boosting algorithm.

The search space comprised seven Gradient Boosting hyperparameters with the candidate values shown in Table 7. RandomizedSearchCV evaluated 60 randomly selected parameter combinations from this search space using 5-fold cross-validation, with the coefficient of determination (R^2) as the scoring criterion. Hyperparameter optimization was performed exclusively on the training dataset, while the testing dataset was reserved for final model evaluation. The testing dataset was not used during hyperparameter optimization and was reserved exclusively for the final evaluation of the selected model. After

the optimal hyperparameter configuration was obtained, Repeated Cross-Validation was conducted on the training dataset to further assess the stability of the final model across repeated data partitions. This validation was conducted separately from the hold-out testing procedure. The repeated cross-validation results were used to assess model stability, whereas the 20% testing set remained reserved for the final evaluation of predictive performance on unseen observations.

Table 7. Hyperparameter Search Space for Gradient Boosting.

Hyperparameter	Values Tested
n_estimators	300, 500, 700, 1000, 1500
learning_rate	0.03, 0.05, 0.07, 0.01, 0.15
max_depth	2, 3, 4, 5, 6
min_samples_split	2, 3, 5, 10
min_samples_leaf	1, 2, 3
subsample	0.70, 0.80, 0.90, 1.00
max_features	sqrt, log2, None

The optimal hyperparameter configuration obtained from the optimization process is presented in Table 8.

Table 8. Optimal Hyperparameters of the Gradient Boosting Model.

Hyperparameter	Optimal Value
Number of Trees (n_estimators)	300
Learning Rate (learning_rate)	0.03
Maximum Tree Depth (max_depth)	3
Subsample Ratio (subsample)	0.70
Minimum Samples Split (min_samples_split)	10
Minimum Samples Leaf (min_samples_leaf)	2
Maximum Features (max_features)	sqrt

The resulting hyperparameter configuration was used as the final setting of the Gradient Boosting model to achieve optimal predictive performance for construction cost overrun prediction. A relatively small learning rate combined with a larger number of decision trees enabled the model to learn gradually and more stably, while limiting tree depth and employing subsampling techniques helped reduce the risk of overfitting and improve the model's generalization capability on unseen observations.

2.7. Model Evaluation Metrics

The performance of the Gradient Boosting model was evaluated using four performance metrics that are commonly adopted in construction cost prediction studies, namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). The combination of these metrics provides a comprehensive assessment of both prediction accuracy and model generalization capability. Mean Absolute Error (MAE) was used to measure the average magnitude of prediction errors without considering their direction. Lower MAE values indicate higher prediction accuracy and were calculated using Equation (2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Root Mean Square Error (RMSE), on the other hand, assigns a larger penalty to extreme prediction errors and is therefore more sensitive to the presence of outliers. RMSE was calculated using Equation (3):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathcal{Y}_i - \hat{\mathcal{Y}}_i)^2} \quad (3)$$

To evaluate the relative prediction error with respect to the observed latent Cost Overrun construct scores, Mean Absolute Percentage Error (MAPE) was employed. This metric expresses prediction errors as percentages and was calculated using Equation (4):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\mathcal{Y}_i - \hat{\mathcal{Y}}_i}{\mathcal{Y}_i}\right| \quad (4)$$

Because the prediction target is a latent construct score rather than an observed monetary or percentage-based cost overrun, MAPE is interpreted as a relative error measure on the latent construct scale. In addition, the coefficient of determination (R^2) was used to evaluate the ability of the model to explain the variability of the observed data. An R^2 value closer to one indicates better predictive capability and was calculated using Equation (5):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\mathcal{Y}_i - \hat{\mathcal{Y}}_i)^2}{\sum_{i=1}^{n}(\mathcal{Y}_i - \overline{\mathcal{Y}}_i)^2} \quad (5)$$

The combined use of MAE, RMSE, MAPE, and R^2 enables a more comprehensive evaluation of model performance because each metric exhibits different sensitivities to prediction errors and model behavior. Furthermore, this evaluation framework has been widely adopted in machine learning-based construction cost prediction studies, thereby facilitating direct comparison of model performance with previous research findings.

3. Results

3.1. Measurement Model Assessment

The measurement model evaluation results indicate that all research constructs satisfied the validity and reliability criteria required in SEM-PLS analysis. The outer loading results are presented in Table 3, while the cross-loading results are presented in Table 4. The HTMT results and VIF assessment are presented in Tables 5 and 6, respectively. As presented in Table 9, all constructs exhibited Cronbach's Alpha and Composite Reliability (CR) values exceeding the minimum threshold of 0.70, indicating a satisfactory level of internal consistency among the indicators forming each construct. Furthermore, all constructs achieved Average Variance Extracted (AVE) values above the recommended threshold of 0.50, thereby satisfying the convergent validity criterion commonly adopted in SEM-PLS studies [32]. These results indicate that the research indicators were able to explain more than 50% of the variance of their respective constructs and can therefore be considered adequate representations of the latent variables included in the analysis.

The Geographical Factor construct demonstrated the highest level of reliability, with a Composite Reliability value of 0.942 and an AVE value of 0.802, indicating excellent indicator homogeneity and explanatory capability. Meanwhile, the Risk Factor construct exhibited the highest Cronbach's Alpha value (0.952), reflecting a very high level of internal consistency among the project risk indicators. Overall, these findings confirm that the SEM-PLS measurement model successfully validated the 24 indicators and established stable and reliable representations of the three theoretical constructs, providing the empirical basis for their subsequent use in the Gradient Boosting model development stage.

Table 9. Measurement Model Assessment.

Construct	Cronbach's Alpha	Composite Reliability	AVE
Social Factor (X1)	0.748	0.834	0.558
Geographical Factor (X2)	0.918	0.942	0.802
Risk Factor (X3)	0.952	0.957	0.582
Latent Cost Overrun construct (Y)	0.830	0.898	0.745

3.2. Structural Model Assessment

Following the establishment of construct validity and reliability, the structural model was evaluated to identify the relationships among latent variables and determine the factors influencing cost overruns in construction projects. The analysis results indicated that all independent variables exhibited statistically significant relationships with the latent Cost Overrun construct. The structural model results were assessed using the path coefficient, bootstrapped standard deviation, T-statistic, and *p*-value. As presented in Table 10, all three relationships achieved *p*-values below the significance threshold of 0.05. All 24 predictor indicators were retained for the subsequent Gradient Boosting model development stage based on the overall measurement-model assessment. As presented in Table 10, the Geographical Factor (X2) exhibited the strongest path coefficient with the latent Cost Overrun construct, with a path coefficient of 0.211, a bootstrapped standard deviation of 0.089, a T-statistic of 2.378, and a *p*-value of 0.017. This finding suggests that geographical conditions, regional accessibility, limitations in supporting infrastructure, and logistical challenges contribute significantly to increasing construction project costs in Southwest Papua. The result is consistent with the characteristics of the study area, which is highly dependent on the transportation of materials and equipment from other regions and is therefore particularly vulnerable to supply chain disruptions and transportation costs.

Table 10. Structural Model Assessment.

Relationship	Path Coefficient	Bootstrapped Standard Deviation	T-Statistic	*p*-Value
Social Factor (X1) → Latent Cost Overrun construct (Y)	0.163	0.083	1.969	0.049
Geographical Factor (X2) → Latent Cost Overrun construct (Y)	0.211	0.089	2.378	0.017
Risk Factor (X3) → Latent Cost Overrun construct (Y)	0.067	0.026	2.537	0.011

The Social Factor (X1) represented the second strongest path coefficient with the latent Cost Overrun construct, with a path coefficient of 0.163, a bootstrapped standard deviation of 0.083, a T-statistic of 1.969, and a *p*-value of 0.049. This finding indicates that the interaction between construction projects and local communities, including community acceptance, social conflicts, local labor participation, and adaptation to local socio-cultural conditions, plays an important role in determining the effectiveness of project cost control.

Meanwhile, the Risk Factor (X3) demonstrated the smallest path coefficient (0.067) but remained statistically significant, with a bootstrapped standard deviation of 0.026, a T-statistic of 2.537, and a *p*-value of 0.011. The smaller path coefficient of the Risk Factor does not imply a lower statistical significance because statistical significance is determined by the ratio between the estimated path coefficient and its bootstrapped standard deviation. In this case, the Risk Factor produced the smallest path coefficient (0.067) but also the smallest standard deviation (0.026), resulting in the highest T-statistic (2.537) and the lowest *p*-value (0.011) among the three structural relationships. Therefore, the ordering of path coefficients does not necessarily correspond to the ordering of *p*-values.

The explanatory capability of the structural model was further evaluated using the coefficient of determination (R^2). The analysis resulted in an R^2 value of 0.587 and an adjusted R^2 value of 0.573, indicating that approximately 58.7% of the variance in the latent Cost Overrun construct was explained by the combined relationships with Social Factors, Geographical Factors, and Risk Factors. The remaining variation was attributed to other factors outside the scope of the present study, including market conditions, policy changes, contractual strategies, and project-specific characteristics that were not incorporated into the research model. The structural model quality indicators are presented in Table 11.

Table 11. Structural Model Quality Indicators.

Indicator	Value
R^2	0.587
Adjusted R^2	0.573
SRMR	0.094

The SRMR value was 0.094. This value is above the commonly referenced threshold of 0.08 and should therefore be interpreted as indicating a marginal level of approximate model fit rather than excellent fit. The SRMR result is reported transparently as a limitation of the overall model fit assessment, while the structural model's explanatory capability is evaluated separately using R^2 and adjusted R^2.

These findings indicate that the SEM-PLS approach provided empirical evidence regarding the validity and relevance of the research constructs and their indicators for the subsequent Gradient Boosting prediction stage. The validated indicator-level scores were retained as machine learning input features, while the three constructs continued to serve as the theoretical framework for organizing these indicator-level features. The SEM-PLS results therefore provided empirical evidence for the relevance of the research constructs and their indicators before the predictive modelling stage.

3.3. Gradient Boosting Model Performance

The performance of the Gradient Boosting model was evaluated using four primary evaluation metrics, namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These metrics were employed to provide a comprehensive assessment of both prediction accuracy and the model's capability to explain variations in the observed data. The evaluation results indicate that the Gradient Boosting model achieved satisfactory predictive performance for the latent Cost Overrun construct, with an MAE value of 1.7838 and an RMSE value of 2.2079. These prediction errors indicate the model's ability to estimate the observed scores of the latent Cost Overrun construct from the validated indicator-level input features obtained through SEM-PLS.

Furthermore, the model produced a MAPE value of 19.10%, indicating the average relative prediction error with respect to the observed latent Cost Overrun construct scores. The Gradient Boosting model achieved an R^2 value of 0.6173, indicating that approximately 61.73% of the variance in the latent Cost Overrun construct scores in the held-out testing set was explained by the validated indicator-level input features. The corresponding SEM-PLS R^2 of 0.587 and adjusted R^2 of 0.573 describe the explanatory capability of the structural relationships among the latent constructs and are therefore reported as a complementary structural result rather than as a benchmark for predictive superiority. Because the predictors and target were derived from the same 233 questionnaire observations and the same SEM-PLS measurement framework, the difference between the two R^2 values should not

be interpreted as statistical evidence that Gradient Boosting outperforms SEM-PLS. The complete performance evaluation results are presented in Table 12.

Table 12. Performance Evaluation of the Gradient Boosting Model for the Latent Cost Overrun Construct.

Metric	Value
MAE	1.7838
RMSE	2.2079
MAPE (%)	19.0962
R^2	0.6173

The results indicate that the Gradient Boosting algorithm can model nonlinear relationships and complex interactions among the validated indicator-level input features that are not explicitly represented in the linear structural relationships assessed through SEM-PLS. This finding supports the application of ensemble learning methods for predicting the latent Cost Overrun construct, particularly in construction projects located in regions characterized by high geographical complexity, social challenges, and project risks, such as Southwest Papua.

Figure 3 illustrates the relationship between the observed scores of the latent Cost Overrun construct and the scores predicted by the Gradient Boosting model on the held-out testing set. The results should be interpreted as prediction performance for the latent Cost Overrun construct rather than as direct prediction of monetary or percentage-based project cost overrun. Most observations are distributed close to the 45° reference line, indicating a high degree of agreement between the observed and predicted construct scores. These findings are consistent with the obtained performance metrics, namely an MAE of 1.7838, an RMSE of 2.2079, a MAPE of 19.10%, and an R^2 value of 0.6173.

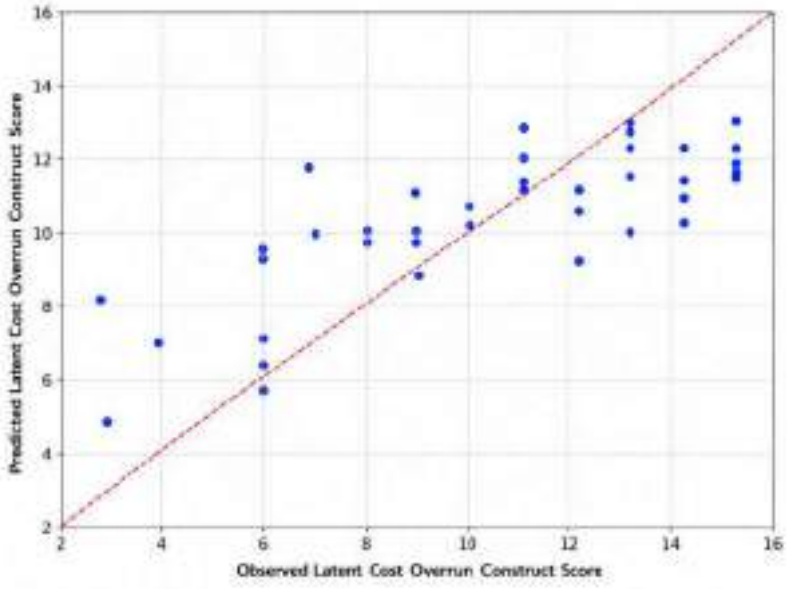


Figure 3. Observed versus predicted scores of the latent Cost Overrun construct on the held-out testing set.

3.4. Comparative Analysis with Other Models

To evaluate the effectiveness of the proposed approach in predicting the latent Cost Overrun construct, a comparative analysis was conducted among the Gradient Boosting, Random Forest, and Artificial Neural Network (ANN) algorithms using four evaluation

metrics, namely MAE, RMSE, MAPE, and R^2. These algorithms were selected because they are among the machine learning approaches commonly applied in construction cost prediction and project management studies.

As shown in Table 13, the Gradient Boosting model achieved the best hold-out test-set performance across the four evaluation metrics considered in this study. The model produced the lowest MAE value (1.7838) and the lowest RMSE value (2.2079) on the hold-out testing set, indicating lower prediction errors than the alternative models in this particular test split. Furthermore, the MAPE value of 19.10% represents the average relative prediction error with respect to the observed scores of the latent Cost Overrun construct, which range from 3 to 15. Therefore, this value should be interpreted as a relative error measure on the constructed latent-score scale rather than as a direct indicator of monetary cost prediction accuracy.

Table 13. Comparative Performance of Machine Learning Algorithms for Predicting the Latent Cost Overrun Construct.

Metric	Random Forest	Gradient Boosting	Artificial Neural Network
MAE	1.8899	1.7838	2.4922
RMSE	2.2266	2.2079	3.0278
MAPE (%)	20.8360	19.0962	26.1725
R^2	0.6107	0.6173	0.2802

From the perspective of explanatory capability on the hold-out testing set, Gradient Boosting achieved the highest R^2 value (0.6173), compared with 0.6107 for Random Forest. However, the difference in R^2 between the two models was small ($\Delta R^2 = 0.0066$), so the observed advantage of Gradient Boosting cannot be interpreted as evidence of statistically significant superiority. Although the difference between the two models was relatively small, the hold-out testing results indicate that Gradient Boosting provided slightly better predictive performance than Random Forest on the predefined test split. This result should be interpreted as a dataset-specific predictive outcome rather than evidence that the boosting mechanism is statistically superior to the bagging approach.

To further assess model stability beyond the single hold-out test split, the Repeated Cross-Validation results presented in Table 14 were also considered. Random Forest achieved a mean MAE of 2.1843 ± 0.2656, RMSE of 2.8022 ± 0.3281, and R^2 of 0.3237 ± 0.1392, whereas Gradient Boosting achieved 2.2305 ± 0.2215, 2.8661 ± 0.2911, and 0.2870 ± 0.1629, respectively. The ANN model produced 2.6284 ± 0.3312 for MAE, 3.3595 ± 0.4178 for RMSE, and -0.0074 ± 0.2841 for R^2. These results indicate that Random Forest exhibited slightly better average performance and stability under repeated cross-validation, while Gradient Boosting remained competitive. Therefore, the selection of Gradient Boosting as the primary model is based on its performance on the predefined hold-out testing set rather than a claim of statistically significant superiority over Random Forest.

Table 14. Repeated Cross-Validation Performance of Machine Learning Models.

Algorithm	MAE (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)
Random Forest	2.1843 ± 0.2656	2.8022 ± 0.3281	0.3237 ± 0.1392
Gradient Boosting	2.2305 ± 0.2215	2.8661 ± 0.2911	0.2870 ± 0.1629
Artificial Neural Network	2.6284 ± 0.3312	3.3595 ± 0.4178	-0.0074 ± 0.2841

In contrast, the ANN model demonstrated relatively inferior performance across all evaluation metrics, particularly in terms of explanatory capability, with an R^2 value of

only 0.2802. These findings indicate that, for datasets with relatively limited sample sizes, ensemble learning approaches such as Gradient Boosting and Random Forest exhibit better generalization capability than artificial neural network models, which typically require larger training datasets to achieve optimal performance.

Overall, the comparative analysis indicates that Gradient Boosting provided the strongest performance on the predefined hold-out testing set, while the repeated cross-validation results showed that Random Forest had slightly better average stability across repeated data partitions. Considering these complementary findings, Gradient Boosting was retained as the primary prediction model because it achieved the lowest MAE and RMSE and the highest R^2 on the predefined hold-out testing set.

3.5. Feature Importance Analysis

Feature importance analysis was performed to identify the relative contribution of the input features to the cost overrun predictions generated by the Gradient Boosting model. In addition to improving model interpretability, this approach supports the principles of Explainable Artificial Intelligence (XAI), which has increasingly been adopted in machine learning-based prediction systems in infrastructure and transportation applications [33,34]. Unlike SEM-PLS results, which describe relationships among theoretical constructs through the structural model, feature importance analysis provides insights into the predictive contribution of individual input features in capturing nonlinear relationships within construction project data. In this study, feature importance was examined at the indicator level, while the corresponding indicators were conceptually grouped into the Social Factor (X1), Geographical Factor (X2), and Risk Factor (X3) constructs.

As illustrated in Figure 4, the Risk Factor (X3) was identified as the most influential variable in predicting Cost Overrun, contributing 53.64% to the overall model prediction, followed by the Social Factor (X1) with 26.03%, and the Geographical Factor (X2) with 20.33%. This finding is consistent with recent studies indicating that risk management has become a critical component in the application of artificial intelligence within modern infrastructure projects [35].

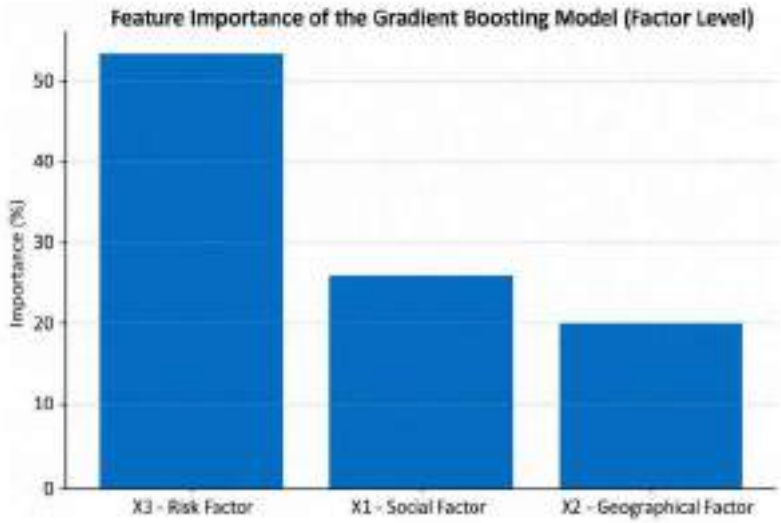


Figure 4. Factor-level feature importance of the Gradient Boosting model.

These results indicate that project risks play a more dominant role in construction cost overruns than the other factors considered in this study. Logistics risks, material price fluctuations, design changes, delays in decision-making processes, and uncertainties

during project implementation were identified as key contributors influencing the model's predictive capability.

Interestingly, these findings differ from the SEM-PLS results, which identified the Geographical Factor as having the largest path coefficient influencing Cost Overrun. This discrepancy highlights the distinction between linear causal relationships identified through SEM-PLS and nonlinear predictive contributions captured by the Gradient Boosting algorithm. The findings suggest that although geographical conditions exert a strong direct influence on cost overruns, project risks exhibit more complex interactions and contribute more substantially to the predictive capability of the model.

At the indicator level, several variables demonstrated relatively stronger contributions than others. The five indicators with the highest importance values are presented in Table 15.

The dominance of these indicators suggests that social aspects, regional accessibility, and operational risk factors contribute significantly to the formation of cost overrun patterns in construction projects in Southwest Papua. These findings have important practical implications, indicating that cost overrun mitigation strategies should not focus solely on technical project aspects but should also consider social conditions and local contextual factors that influence project implementation.

Table 15. Top Five Most Important Indicators in the Gradient Boosting Model.

Rank	Indicator	Importance (%)
1	X1.1	12.48
2	X2.1	7.78
3	X3.5	6.07
4	X1.4	5.47
5	X2.2	5.29

3.6. Discussion and Practical Implications

The results of this study demonstrate that the hybrid approach integrating SEM-PLS and Gradient Boosting provides a complementary framework in which SEM-PLS establishes empirical validity before the predictive modelling stage, while Gradient Boosting is used to capture nonlinear relationships and complex interactions among the validated input features. The differences observed between the SEM-PLS results and the feature importance analysis highlight the distinct characteristics of causal and predictive approaches. While the SEM-PLS analysis identified the Geographical Factor as having the strongest direct effect on Cost Overrun, the Gradient Boosting model identified the Risk Factor as the most influential predictor. These findings suggest that although geographical conditions exert a substantial direct influence on project costs, project risks arising during construction implementation exhibit more complex interactions and play a more dominant role in explaining variations in cost overruns.

A methodological consideration concerns the relationship between the SEM-PLS and Gradient Boosting stages. Both the predictor features and the latent Cost Overrun target were derived from the same 233 questionnaire observations and the same SEM-PLS measurement framework. Consequently, the Gradient Boosting results should be interpreted as an out-of-sample hold-out prediction within the available survey dataset rather than as an independent external validation of the SEM-PLS model. The higher R^2 obtained by Gradient Boosting should therefore not be interpreted as evidence of statistical superiority over SEM-PLS. Instead, the two approaches serve different analytical purposes: SEM-PLS evaluates the structural relationships among theoretical constructs, whereas Gradient Boosting evaluates the predictive contribution of the validated indicator-level latent scores and their nonlinear interactions. Independent validation using external project-

level observations would be required to establish the generalizability of the predictive model beyond the present survey dataset.

These findings have important implications for construction project management, particularly in regions characterized by challenging geographical and logistical conditions such as Southwest Papua. Cost control strategies should not focus solely on technical estimation and quantity control but should also consider project risks, uncertainty in material distribution, changing site conditions, and social dynamics that may significantly influence project implementation. Furthermore, the results indicate that Gradient Boosting provided competitive predictive performance compared with Random Forest and Artificial Neural Networks, with the strongest performance observed on the predefined hold-out testing set. This finding suggests that Gradient Boosting represents an effective alternative for the development of decision support systems aimed at project planning and cost control, especially in situations where historical project data remain limited.

From an academic perspective, this study contributes methodologically through the integration of SEM-PLS and machine learning techniques for construction cost overrun prediction. This approach achieves a balance between model interpretability and predictive accuracy, resulting in a prediction framework that is not only accurate but also theoretically grounded in explaining relationships among research variables. From a practical perspective, the proposed model can be utilized by government agencies, consultants, and contractors as a decision support tool for cost estimation, dominant risk identification, and the development of cost overrun mitigation strategies during the project planning stage. Consequently, the implementation of the proposed model has the potential to improve budget management effectiveness and support infrastructure development in regions characterized by high uncertainty and project complexity.

Beyond serving as a predictive model, the proposed approach is illustrated through a prototype user interface for a potential decision support system to support construction cost estimation processes [36]. Through such a system, users can enter project information and stakeholder assessment inputs and obtain a predicted latent Cost Overrun score together with information regarding the dominant factors influencing the prediction results. This capability may assist government agencies, consultants, and contractors in risk identification and decision-making processes during both the planning and project control stages.

For project promoters or clients, the framework can support early identification of the dominant factors associated with the latent Cost Overrun construct and help prioritize risk mitigation and cost-control actions during project planning and project monitoring. For general contractors, the same information can be used to focus operational responses on the dominant risk factors affecting project execution, including logistical, site-related, social, and other project risks identified by the model. Thus, the framework provides different decision perspectives for clients and contractors while using the same predictive information: clients can use the results to strengthen planning, budget oversight, and risk prioritization, whereas contractors can use them to prioritize operational mitigation and project-control actions.

More broadly, predictive planning based on the proposed framework can support the interests of multiple stakeholders throughout project implementation. For project personnel, early identification of dominant risk and contextual factors can support more informed planning of project activities and resource requirements. For local communities, consideration of social factors such as community acceptance, social conflicts, local labor participation, and adaptation to local socio-cultural conditions can support earlier attention to issues that may affect project implementation. For investors, the predicted Cost Overrun score and identified dominant factors can provide additional information for project risk and budget oversight. For contractors, the same predictive information can support the

prioritization of operational mitigation measures and project-control actions. In this way, predictive planning extends the use of cost-overrun prediction beyond cost estimation toward proactive risk awareness and stakeholder-oriented project management. The input dashboard is shown in Figure 5a, and the prediction results based on the entered values are shown in Figure 5b.

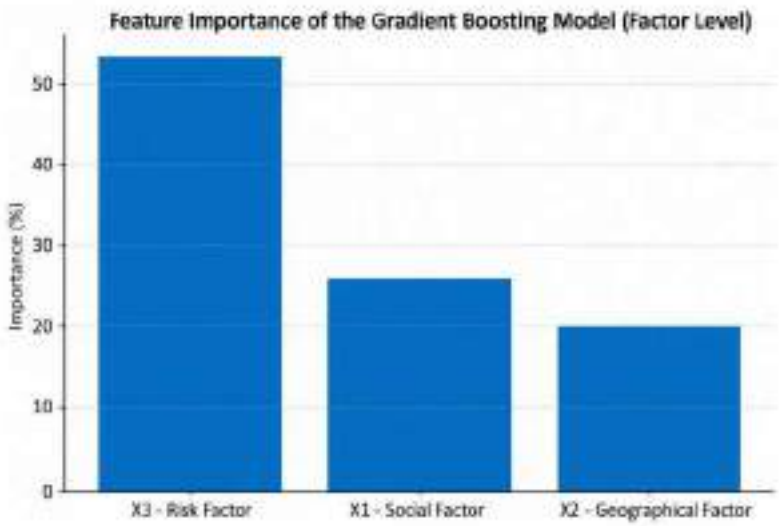


Figure 5. Illustrative Prototype User Interface: (**a**) Input Interface; (**b**) Prediction Output Interface.

4. Conclusions

This study developed a hybrid approach integrating Structural Equation Modeling–Partial Least Squares (SEM-PLS) and Gradient Boosting to predict the latent Cost Overrun construct in construction projects in Southwest Papua Province, Indonesia. The SEM-PLS analysis evaluated the validity and reliability of the 24 indicators representing three primary constructs, namely Social Factors, Geographical Factors, and Risk Factors, and provided the empirical basis for determining the indicators suitable for subsequent predictive modelling. The validated indicator-level latent scores were subsequently used as input features for the Gradient Boosting model. The evaluation results demonstrated that the Gradient Boosting model achieved the strongest predictive performance on the predefined hold-out testing set compared with Random Forest and Artificial Neural Network models, with an MAE of 1.7838, an RMSE of 2.2079, a MAPE of 19.10%, and an R^2 value of 0.6173. These findings indicate that ensemble learning approaches possess a strong capability to capture nonlinear relationships and complex interactions among variables influencing cost overruns in construction projects. The practical interpretation of the proposed model is limited to assessing the level of the latent Cost Overrun construct derived from stakeholder assessments. The model should therefore not be interpreted as a direct estimator of the monetary amount or percentage of actual project cost overrun.

The feature importance analysis revealed that the Risk Factor was the most influential variable in predicting the latent Cost Overrun construct, contributing 53.64% to the overall model prediction, followed by the Social Factor (26.03%) and the Geographical Factor (20.33%). At the indicator level, feature importance was calculated for the validated indicator-level input features used by the Gradient Boosting model. These findings suggest that cost control strategies in construction projects, particularly in regions characterized by challenging geographical conditions, should place greater emphasis on risk management and uncertainty mitigation during project implementation. In addition to its methodological contribution through the integration of SEM-PLS and machine learning, this study also demonstrates practical implementation potential through the development of a prototype

user interface-based decision support system to support construction cost estimation and control processes.

This study has several limitations. First, the prediction target was derived from questionnaire-based latent variable scores rather than observed monetary or percentage-based cost-overrun records. Therefore, the proposed model should be interpreted as predicting the latent Cost Overrun construct and not as directly estimating the monetary magnitude of project cost overruns. Future research should integrate validated historical project cost records with stakeholder-derived indicators to develop and externally validate models for direct monetary or percentage-based cost overrun prediction. Second, the dataset was restricted to construction projects in Southwest Papua Province and did not incorporate additional external factors such as national market dynamics, policy changes, and macroeconomic conditions. Future research should therefore expand the geographical scope of the study, increase the volume of historical project data, incorporate additional external variables, and explore explainable artificial intelligence approaches to further enhance the generalizability and interpretability of the proposed prediction model.

Author Contributions: Conceptualization, J.A. and A.P.I.; methodology, J.A., A.P.I. and M.W.; software, J.A.; validation, A.P.I. and M.W.; formal analysis, J.A.; investigation, J.A.; data curation, J.A.; writing—original draft preparation, J.A.; writing—review and editing, A.P.I. and M.W.; visualization, J.A.; supervision, A.P.I. and M.W.; project administration, A.P.I.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request. The data are not publicly available due to confidentiality agreements with project stakeholders and organizations involved in the study.

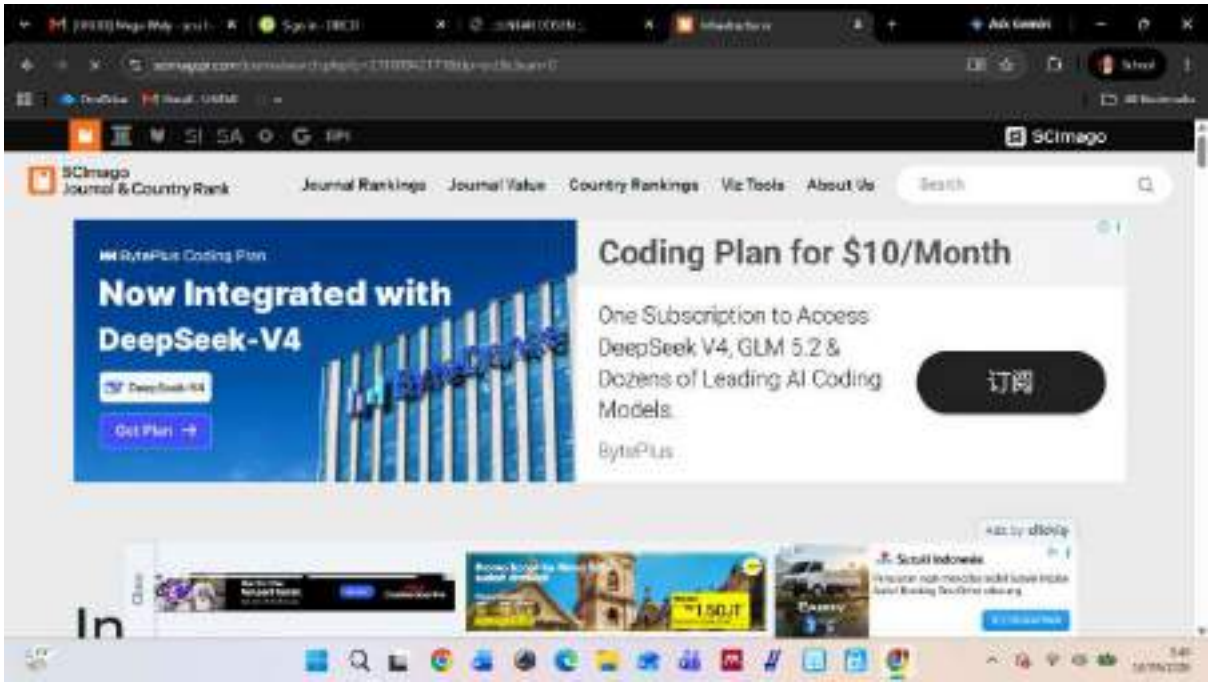
Conflicts of Interest: The authors declare no conflicts of interest.

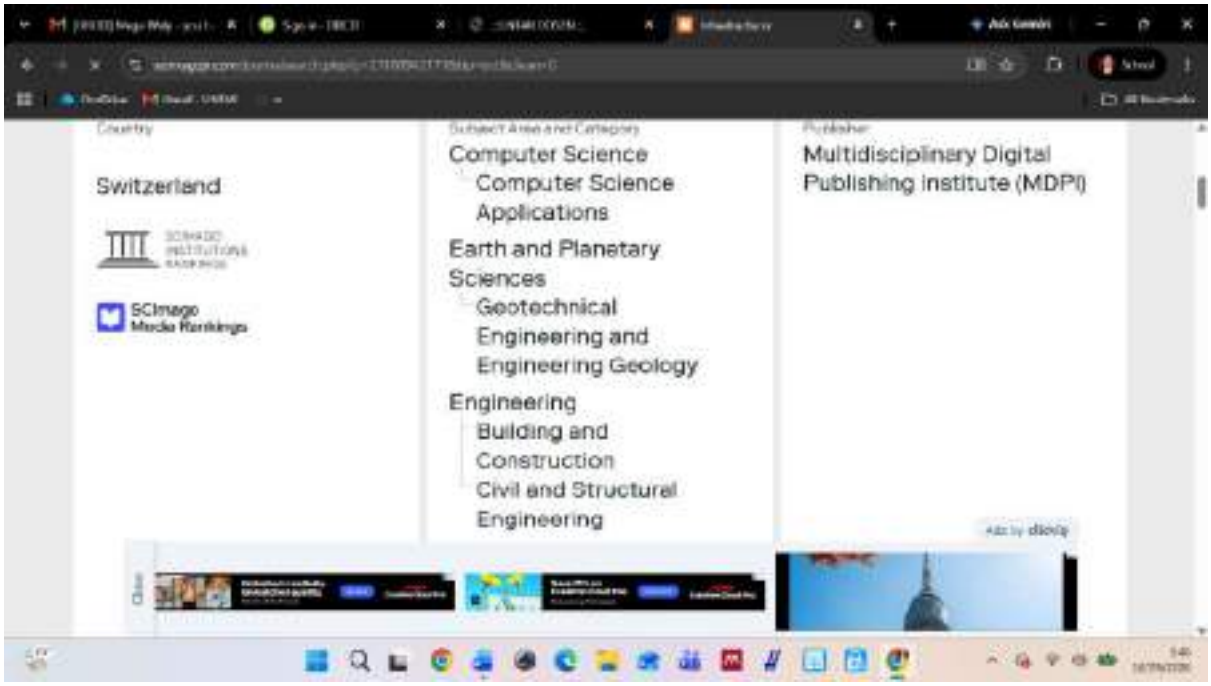
References

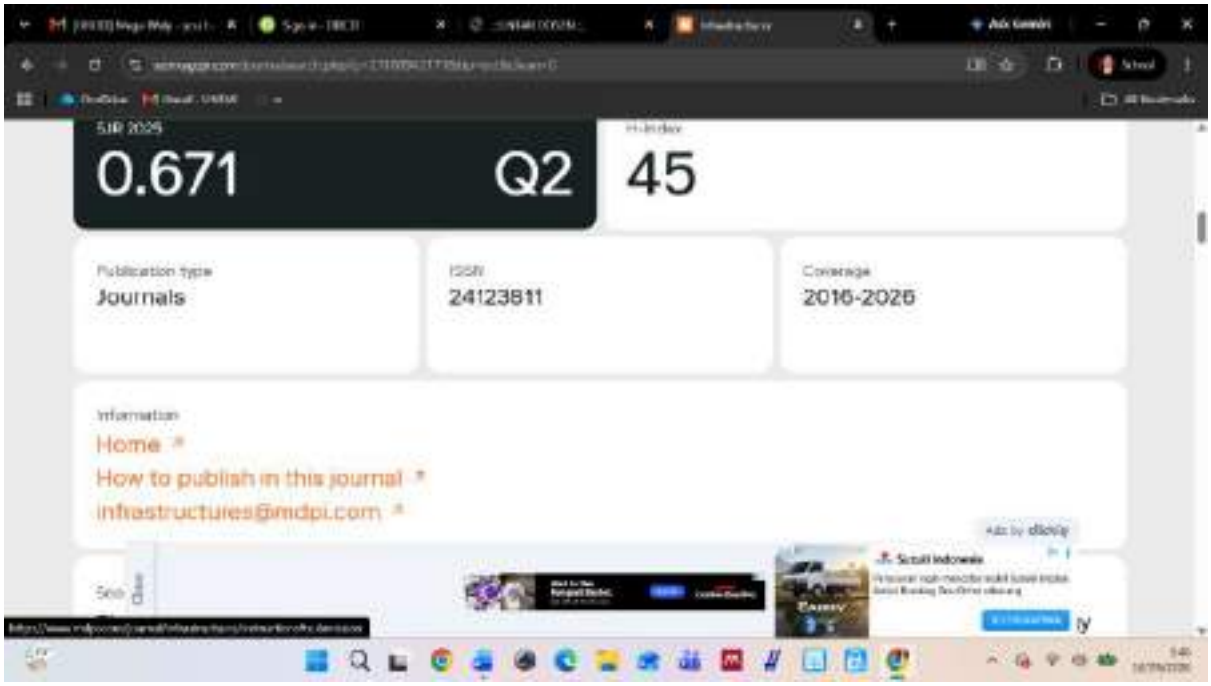
1. Flyvbjerg, B. Survival of the Unfittest: Why the Worst Infrastructure Gets Built—And What We Can Do About It. *Oxf. Rev. Econ. Policy* **2009**, *25*, 344–367. [CrossRef]
2. Love, P.E.D.; Edwards, D.J.; Irani, Z. Moving Beyond Optimism Bias and Strategic Misrepresentation: An Explanation for Social Infrastructure Project Cost Overruns. *IEEE Trans. Eng. Manag.* **2012**, *59*, 560–571. [CrossRef]
3. Cantarelli, C.C.; Van Wee, B.; Molin, E.J.E.; Flyvbjerg, B. Different cost performance: Different determinants? The case of cost overruns in Dutch transport infrastructure projects. *Transp. Policy* **2012**, *22*, 88–95. [CrossRef]
4. Aljohani, A. Construction Projects Cost Overrun: What Does the Literature Tell Us? *Int. J. Innov. Manag. Technol.* **2017**, *8*, 137–143. [CrossRef]
5. Morris, P.W.G.; Jamieson, A.; Shepherd, M.M. Corrigendum to 'Research updating the APM Body of Knowledge 4th edition' [Int. J. Proj. Manag. 24 (2006) 461–473]. *Int. J. Proj. Manag.* **2014**, *32*, 1083. [CrossRef]
6. Dagbui, D.D.A.; Smith, S.D. Dealing with construction cost overruns using data mining. *Constr. Manag. Econ.* **2014**, *32*, 682–694. [CrossRef]
7. Gunduz, M.; Yahya, A.M.A. Analysis of project success factors in construction industry. *Technol. Econ. Dev. Econ.* **2018**, *24*, 67–80. [CrossRef]
8. Doloi, H. Cost Overruns and Failure in Project Management: Understanding the Roles of Key Stakeholders in Construction Projects. *J. Constr. Eng. Manag.* **2014**, *139*, 267–279. [CrossRef]
9. Sundararaman, M.; Sambasivam, B. The Road to Net Zero in a Renewable Energy-Dominated Electricity System: Impact of EV Charging and Social Cost of Emission on the Optimal Economic Dispatch. *Green Energy Intell. Transp.* **2025**, *4*, 1–14. [CrossRef]
10. Chen, Y.; Ahiaga-Dagbui, D.D.; Thaheem, M.J.; Shrestha, A. Toward a Deeper Understanding of Optimism Bias and Transport Project Cost Overrun. *Proj. Manag. J.* **2023**, *54*, 561–578. [CrossRef]
11. Dong, S.; Ahmed, M.; Chatpattananan, V. Analysis of Key Factors of Cost Overrun in Construction Projects Based on Structural Equation Modeling. *Sustainability* **2025**, *17*, 2119. [CrossRef]
12. Abdelalim, A.M.; Salem, M.; Salem, M.; Al-Adwani, M.; Tantawy, M. An Analysis of Factors Contributing to Cost Overruns in the Global Construction Industry. *Buildings* **2025**, *15*, 18. [CrossRef]

13. Abed, Y.G.; Hasan, T.M.; Zehawi, R.N.; Maseer, Z.K. Machine Learning Algorithms for Constructions Cost Prediction: A Systematic Review. *Int. J. Nonlinear Anal. Appl.* **2022**, *13*, 2205–2218. [CrossRef] [PubMed]
14. Das, P.; Kashem, A.; Hasan, I.; Islam, M. A comparative study of machine learning models for construction costs prediction with natural gradient boosting algorithm and SHAP analysis. *Asian J. Civ. Eng.* **2024**, *25*, 3301–3316. [CrossRef]
15. Chakraborty, D.; Elhegazy, H.; Elzarka, H.; Gutierrez, L. A novel construction cost prediction model using hybrid natural and light gradient boosting. *Adv. Eng. Inform.* **2020**, *46*, 101201. [CrossRef]
16. Yan, H.; He, Z.; Gao, C.; Xie, M.; Sheng, H.; Chen, H. Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm. *Adv. Eng. Inform.* **2022**, *54*, 101789. [CrossRef]
17. Sapkota, S.C.; Haq, M.; Thapa, B.; Adhikari, S.; Dhakal, A.; Paudel, R.; Ghimire, A.; Bansal, T. Predicting the Strength of Sustainable Graphene-Enhanced Cementitious Composites Using Novel Machine Learning and Explainable AI Techniques. *Infrastructures* **2026**, *11*, 146. [CrossRef]
18. Mirzaei, L.; Fedler, C.B.; Ghebrab, T. Machine Learning for Relative Compressive Strength of Concrete Incorporating Agricultural Bio-Supplementary Cementitious Materials. *Infrastructures* **2026**, *11*, 190. [CrossRef]
19. Chen, L.; Xu, C.; Lim, W.H.; Sharma, A.; Tiang, S.S.; Chong, K.S.; El-Kenawy, E.-S.M.; Alhussan, A.A.; Eid, M.M.; Khafaga, D.S. Transparent and reliable construction cost prediction using advanced machine learning and explainable AI. *Eng. Sci. Technol. Int. J.* **2025**, *70*, 102159. [CrossRef]
20. Fan, C.-L. Optimization and performance evaluation of machine learning classifiers for predicting construction quality and schedule. *Autom. Constr.* **2025**, *179*, 106470. [CrossRef]
21. Turkyilmaz, A.H.; Polat, G. Risk-Based Completion Cost Overrun Ratio Estimation in Construction Projects Using Machine Learning Classification Algorithms: A Case Study. *Buildings* **2024**, *14*, 3541. [CrossRef]
22. Gao, Y.; Antwi-Afari, M.F.; Huang, Y.; Chen, Z.S.; Manzoor, B. Artificial Intelligence in Construction Project Management: A Systematic Literature Review of Cost, Time, and Safety Management. *Buildings* **2026**, *16*, 1061. [CrossRef]
23. Alazawy, S.F.M.; Ahmed, M.A.; Raheem, S.H.; Imran, H.; Bernardo, L.F.A.; Pinto, H.A.S. Explainable Machine Learning to Predict the Construction Cost of Power Plant Based on Random Forest and Shapley Method. *CivilEng* **2025**, *6*, 21. [CrossRef]
24. Park, U.; Kang, Y.; Lee, H.; Yun, S. A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs. *Appl. Sci.* **2022**, *12*, 9729. [CrossRef]
25. Sadikoglu, E.; Demirkesen, S. Mapping the state-of-the-art: Bibliometric and systematic analysis of machine learning and AI-based cost and time prediction in construction. *J. Asian Archit. Build. Eng.* **2026**, *25*, 1–28. [CrossRef]
26. Zhang, W. Applications of machine learning and artificial intelligence in construction project cost prediction: A scientometric analysis and qualitative review. *Front. Built Environ.* **2026**, *12*, 1867673. [CrossRef]
27. ForouzeshNejad, A.A.; Arabikhan, F.; Aheleroff, S. Optimizing Project Time and Cost Prediction Using a Hybrid XGBoost and Simulated Annealing Algorithm. *Machines* **2024**, *12*, 867. [CrossRef]
28. Friedman, J.H. 999 Reitz Lecture Greedy Function Approximation: A Gradient Boosting Machine 1. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
29. Alshboul, O.; Shehadeh, A.; Almasabha, G.; Almuflih, A.S. Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction. *Sustainability* **2022**, *14*, 6651. [CrossRef]
30. Coffie, G.H.; Cudjoe, S.K.F. Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. *Int. J. Constr. Manag.* **2024**, *24*, 1742–1750. [CrossRef]
31. Setiawan, D.; Putranto, L.S.; Sari, E.M. Application of XGBoost in Road Maintenance Cost Prediction. *Civ. Eng. J.* **2026**, *12*, 1548–1563. [CrossRef]
32. Hair, J.F.; Hult, G.T.M.; Ringle, C.M.; Sarstedt, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 3rd ed.; SAGE Publications: Los Angeles, CA, USA, 2022.
33. Hu, Y.; Atta, Z.; Rahman, T.U.; Qiu, S.; Wang, J.; Wei, W.; Liang, Z.; Zaheer, Q. Shedding Light on Explainable AI: Insights, Challenges, and the Future of Infrastructure Management. *ISPRS Int. J. Geoinf.* **2026**, *15*, 100. [CrossRef]
34. Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
35. Chang, C.M.; Hossain, A. A Climate Adaptation Asset Risk Management Approach for Resilient Roadway Infrastructure. *Infrastructures* **2024**, *9*, 226. [CrossRef]
36. Almulhim, A.I. Integrating Artificial Intelligence into Smart Infrastructure Management for Sustainable Urban Planning. *Technologies* **2025**, *13*, 481. [CrossRef]

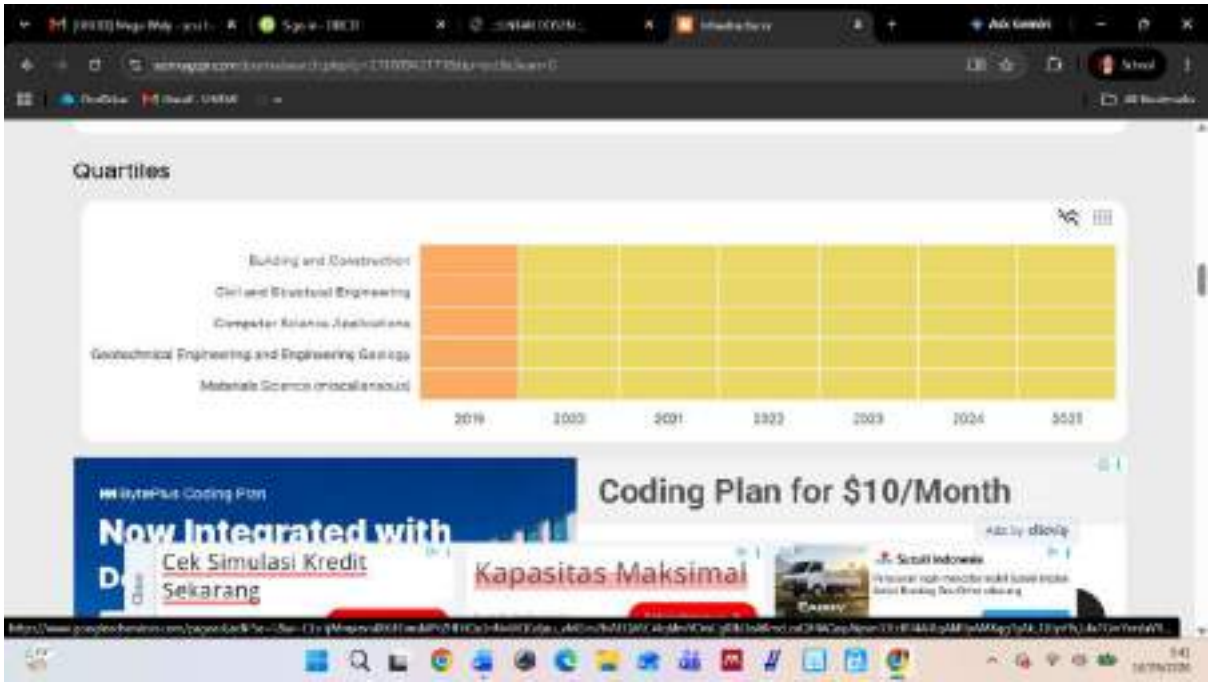
Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

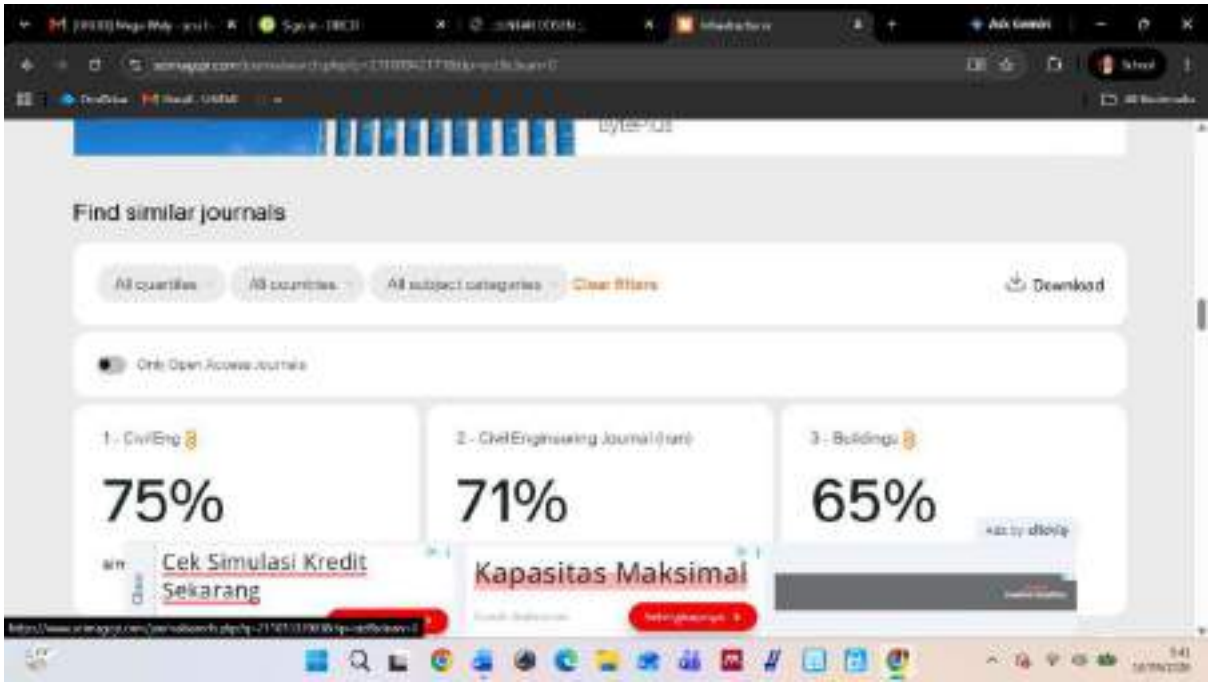

SCImago
SCImago Journal & Country Rank
Journal Rankings
Journal Value
Country Rankings
Viz Tools
About Us
Search
BytePlus Coding Plan
Now Integrated with DeepSeek-V4
Get Plan
Coding Plan for $10/Month
One Subscription to Access DeepSeek V4, GLM 5.2 & Dozens of Leading AI Coding Models
BytePlus
订阅


Country
Switzerland
SCImago Media Rankings
Subject Area and Category
Computer Science
Computer Science Applications
Earth and Planetary Sciences
Geotechnical Engineering and Engineering Geology
Engineering
Building and Construction
Civil and Structural Engineering
Publisher
Multidisciplinary Digital Publishing Institute (MDPI)

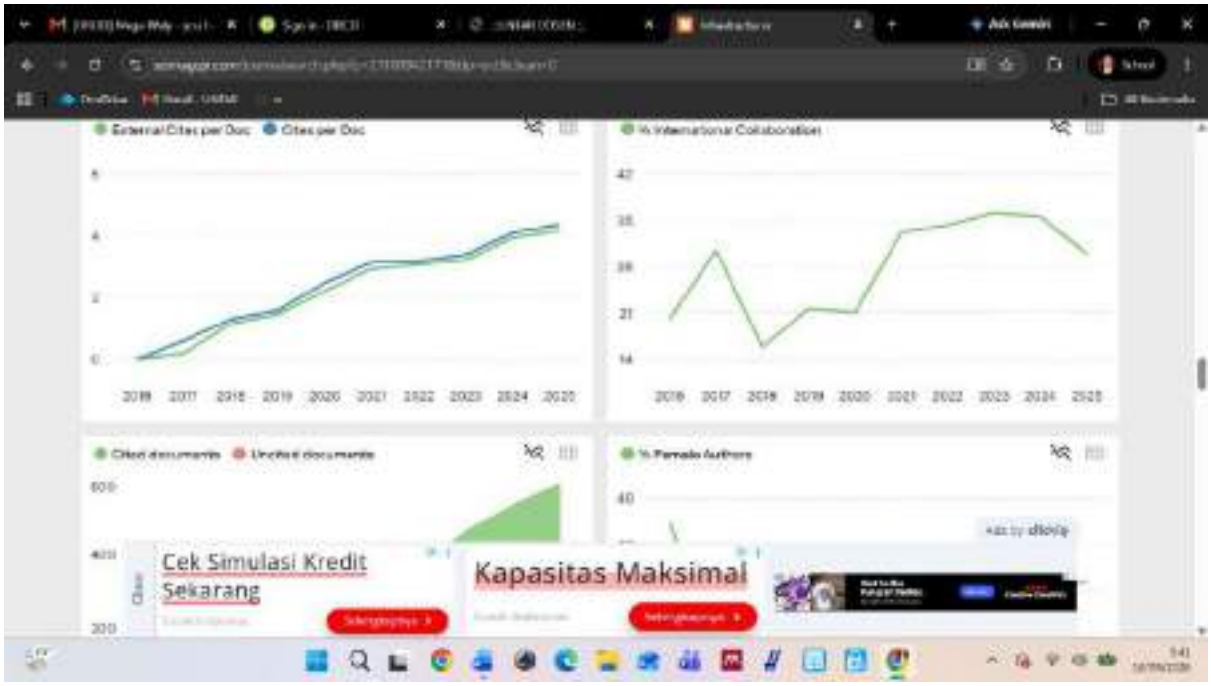

SJR 2025
0.671
Q2
45
Publication type
Journals
ISSN
24123811
Coverage
2016-2025
Information
Home
How to publish in this journal
infrastructures@mdpi.com

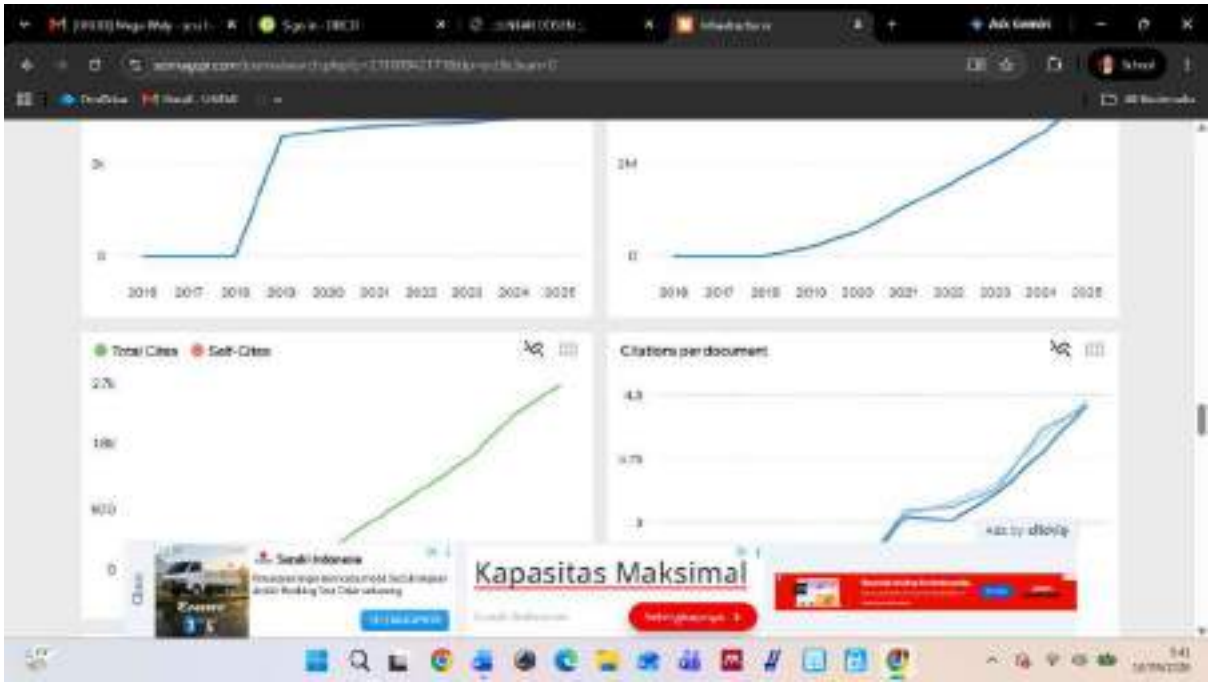
Scope
The scope of Infrastructures includes applications in road, railway, coastal, airport, and energy infrastructures, and encompasses the following: • Construction technology, maintenance, retrofitting, and management of constructed facilities. • Construction materials. • Analysis of transportation networks, logistics, supply chains, management of transport systems, intelligent transportation systems, safety, mobility, and environmental concerns. • Structural design, structural control, health monitoring and prediction, non-destructive technologies, inspection, structural reliability / stability, life cycle assessment and updating, optimization, and integrity. • Pavement engineering. • Railway engineering. • Bridge engineering. • Tunnel engineering. • Dam engineering. • Port and marine engineering. • Geotechnical engineering. • Survey engineering. • Computers in civil engineering. • Smart infrastructures. • Sustainable infrastructures. • Aging infrastructures. • Adaptation to climate change, extreme events, and hazards and hazards mitigation. • Financial planning.

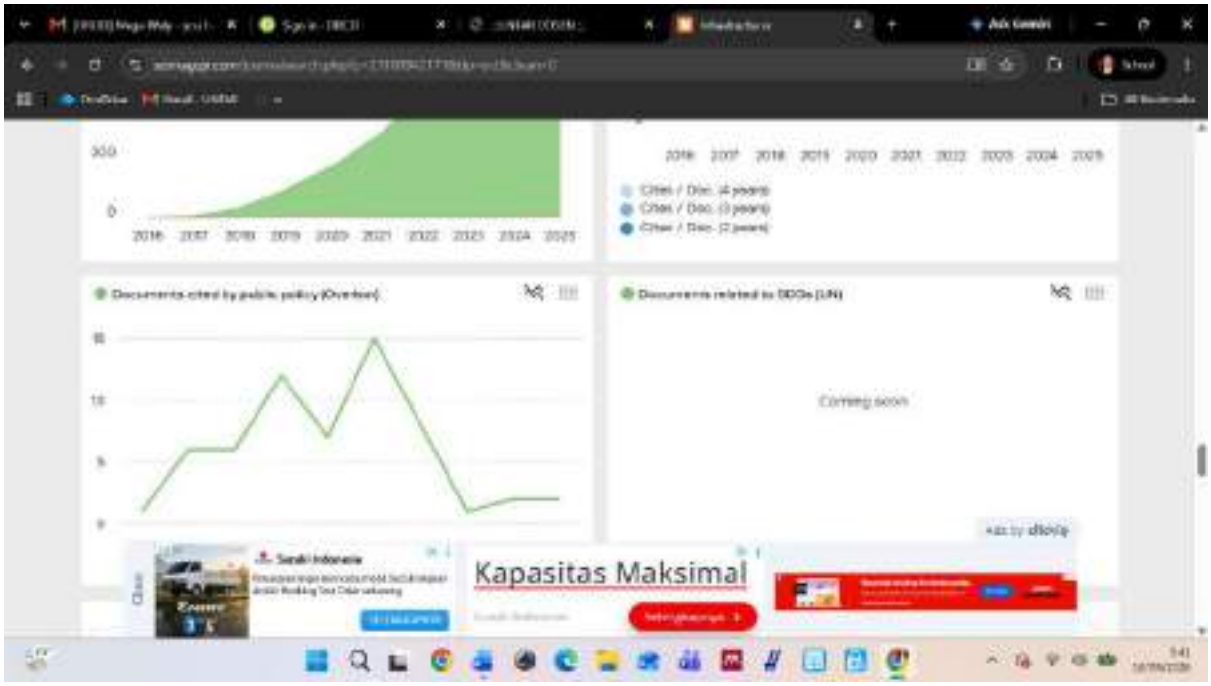

Quartiles
Building and Construction
Civil and Structural Engineering
Geotechnical Engineering and Engineering Geology
2019
2020
2021
2022
2023
2024
2025
Coding Plan for $10/Month
Now Integrated with
Cek Simulasi Kredit Sekarang
Kapasitas Maksimal

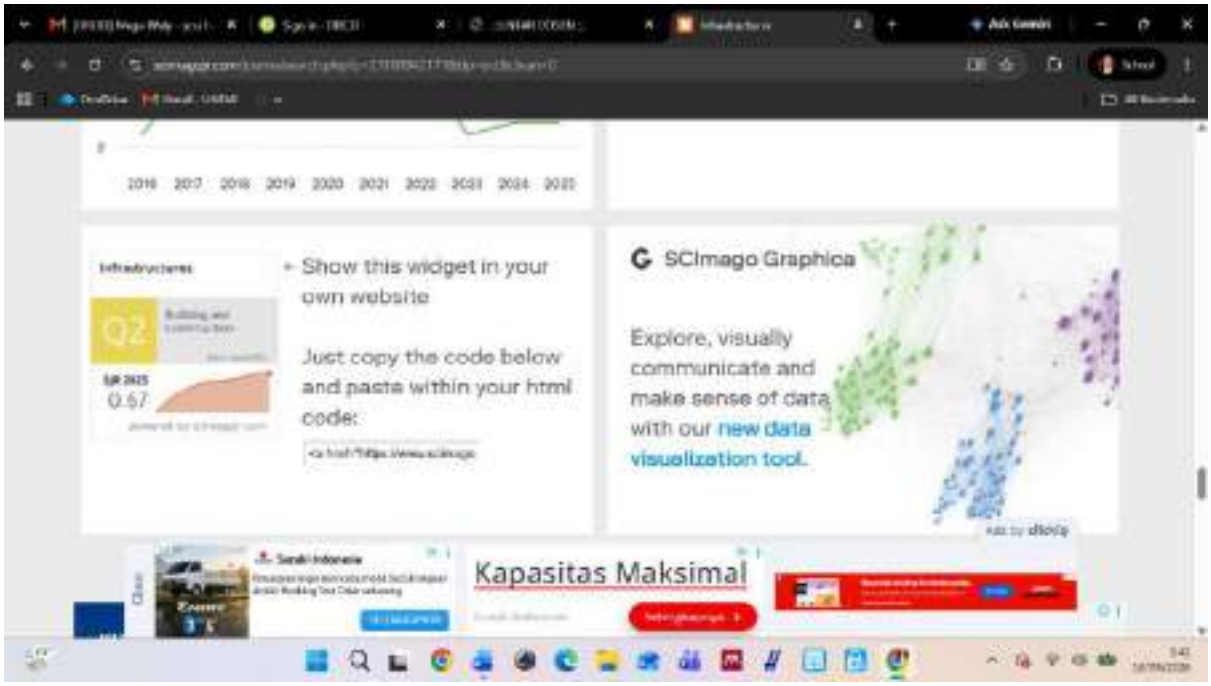

Find similar journals
All quartiles
All countries
All subject categories
Clear filters
Download
Only Open Access Journals
1 - CivilEng
75%
2 - CivilEngineering Journal (Iran)
71%
3 - Buildings
65%
Cek Simulasi Kredit Sekarang
Kapasitas Maksimal

SJR
Total Documents
External Cites per Doc
Cites per Doc
% International Collaboration
Cek Simulasi Kredit Sekarang
Kapasitas Maksimal


External Cites per Doc
Cites per Doc
% International Collaboration
Cited documents
Uncited documents
% Female Authors
Cek Simulasi Kredit Sekarang
Kapasitas Maksimal

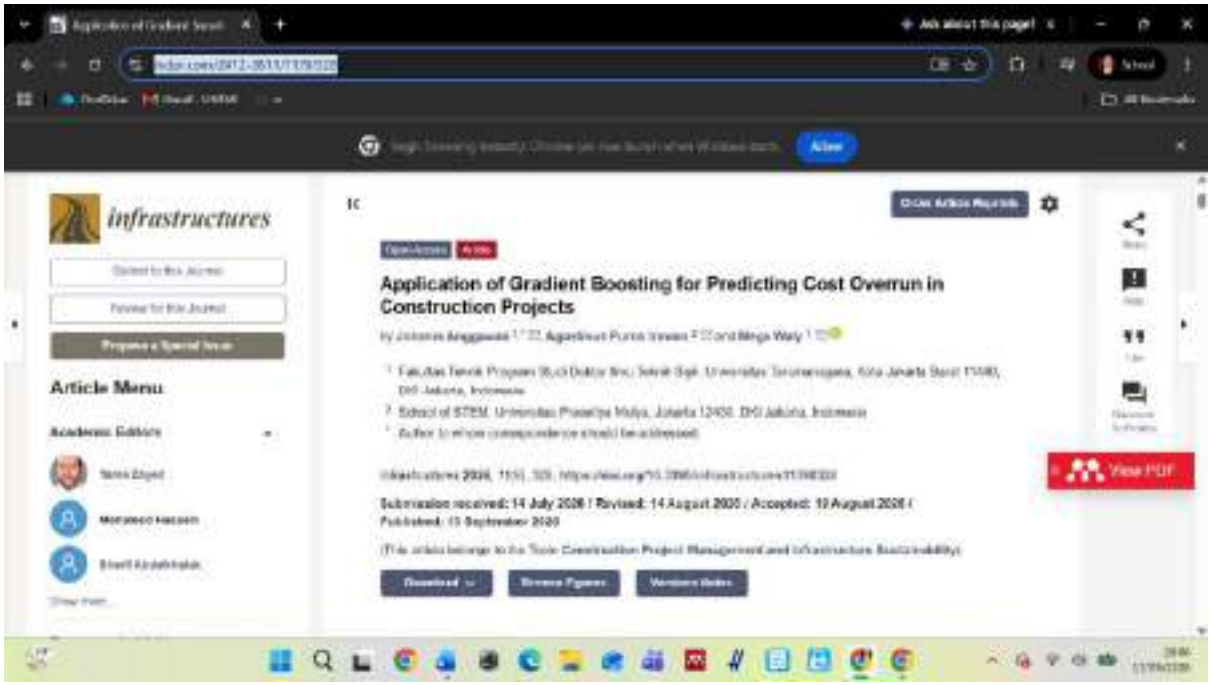

Total Cites
Self-Cites
Citations per document
Kapasitas Maksimal


Documents cited by public policy (Overton)
Documents related to SDGs (UN)
Coming soon
Kapasitas Maksimal


2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
Infrastructures
Q2
SJR 2025
0.67
Show this widget in your own website
Just copy the code below and paste within your html code:
SCImago Graphica
Explore, visually communicate and make sense of data with our new data visualization tool.
Kapasitas Maksimal

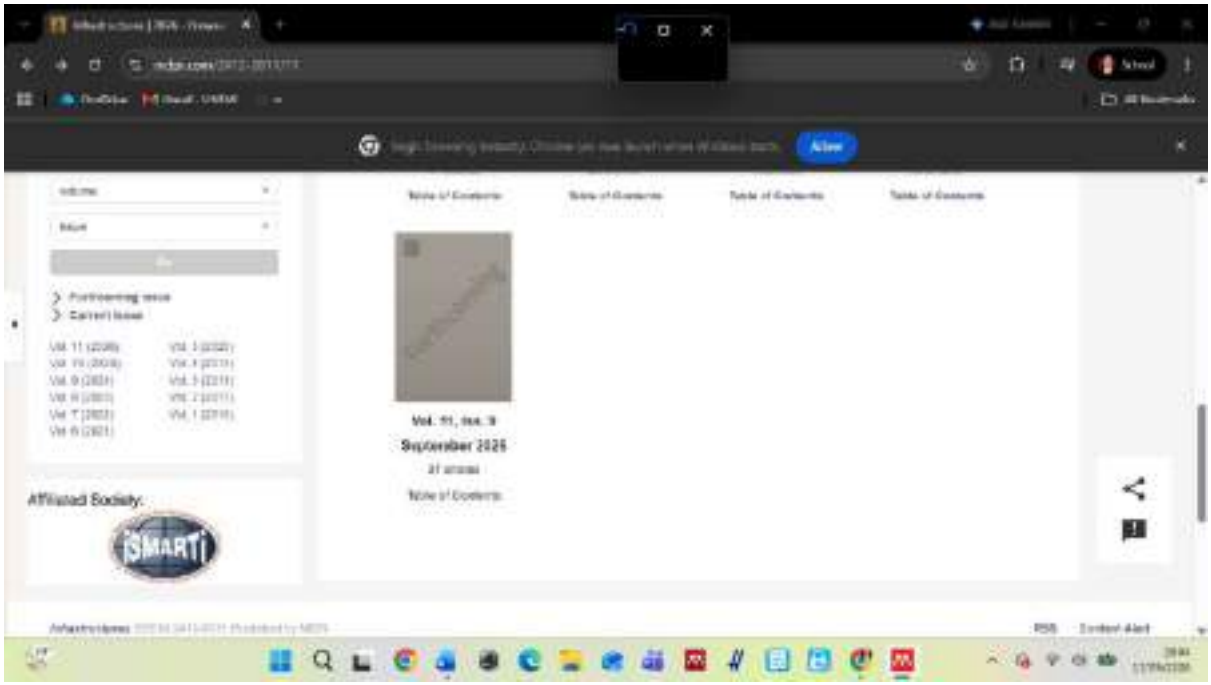
Url Jurnal

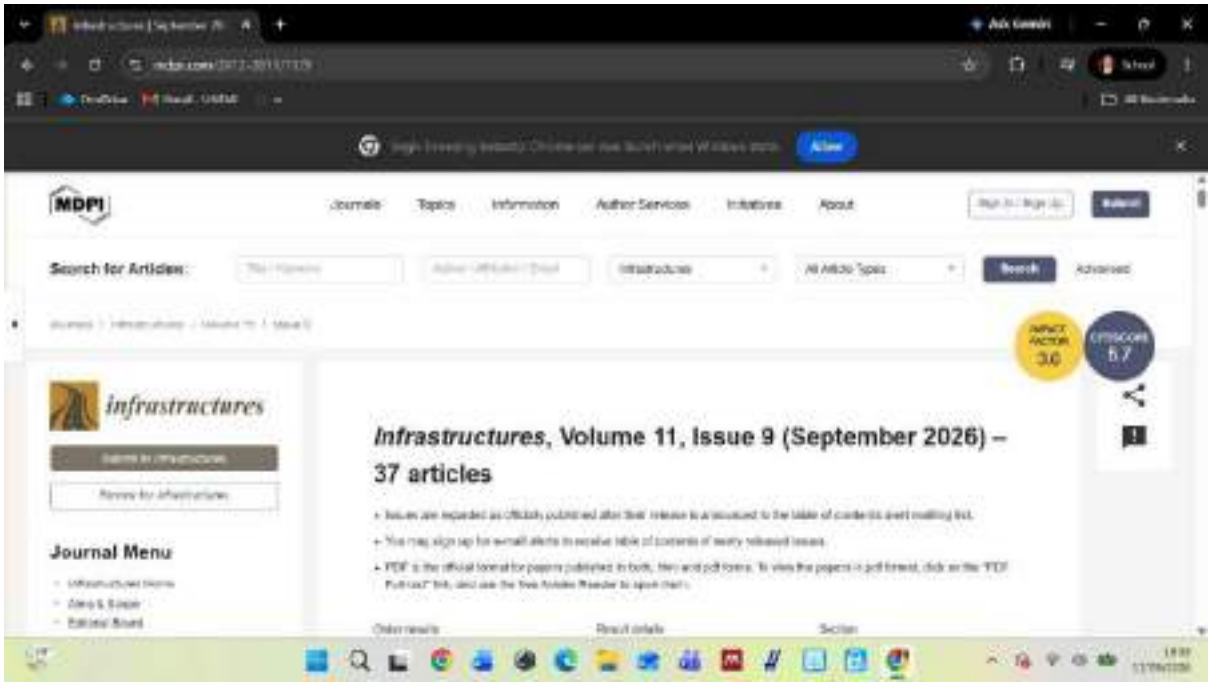
https://www.mdpi.com/2412-3811/11/9/328

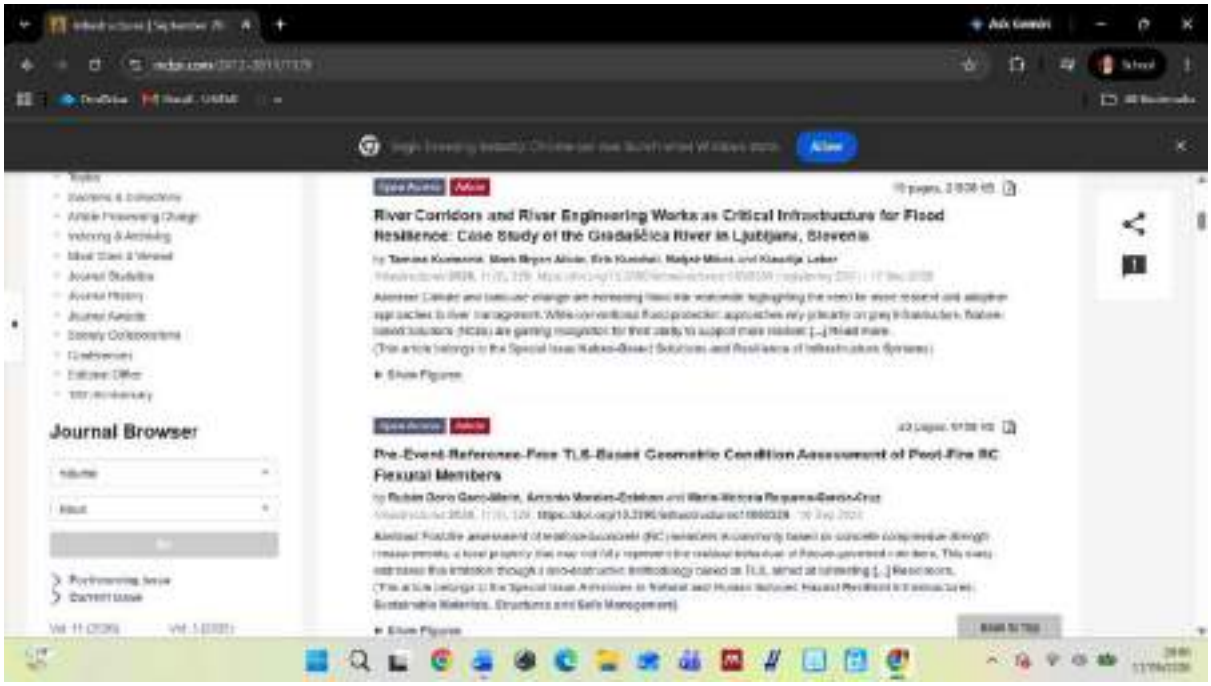


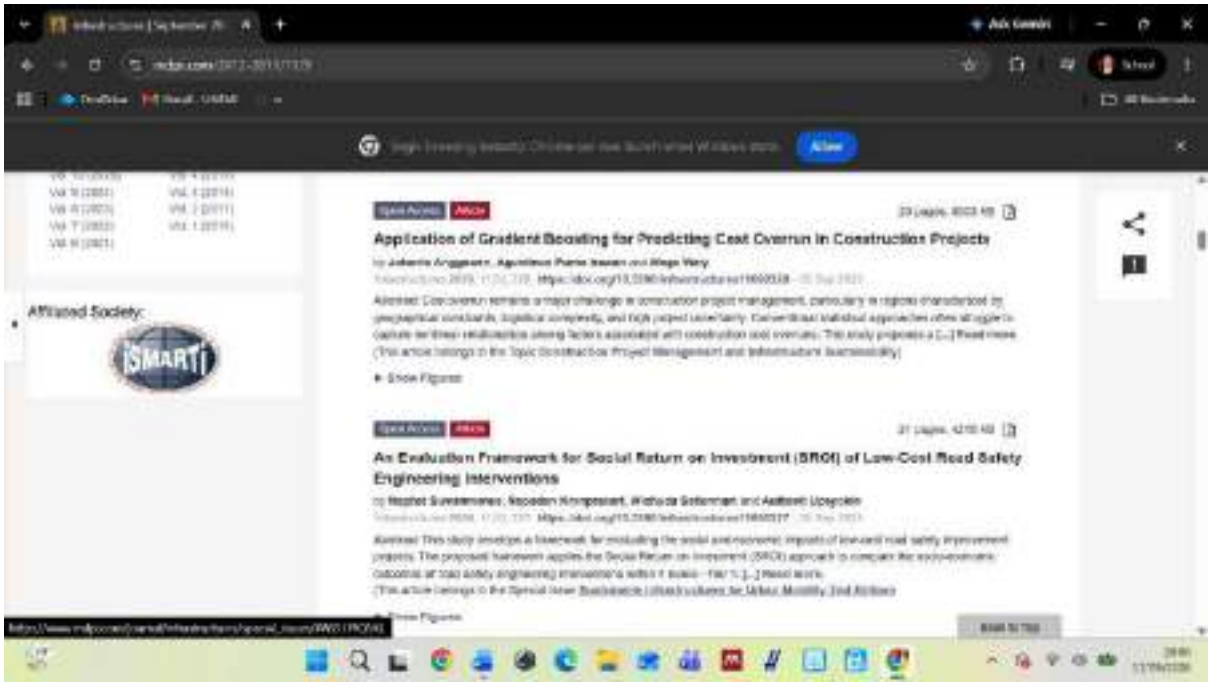
Cover dan daftar isi Infrastructures

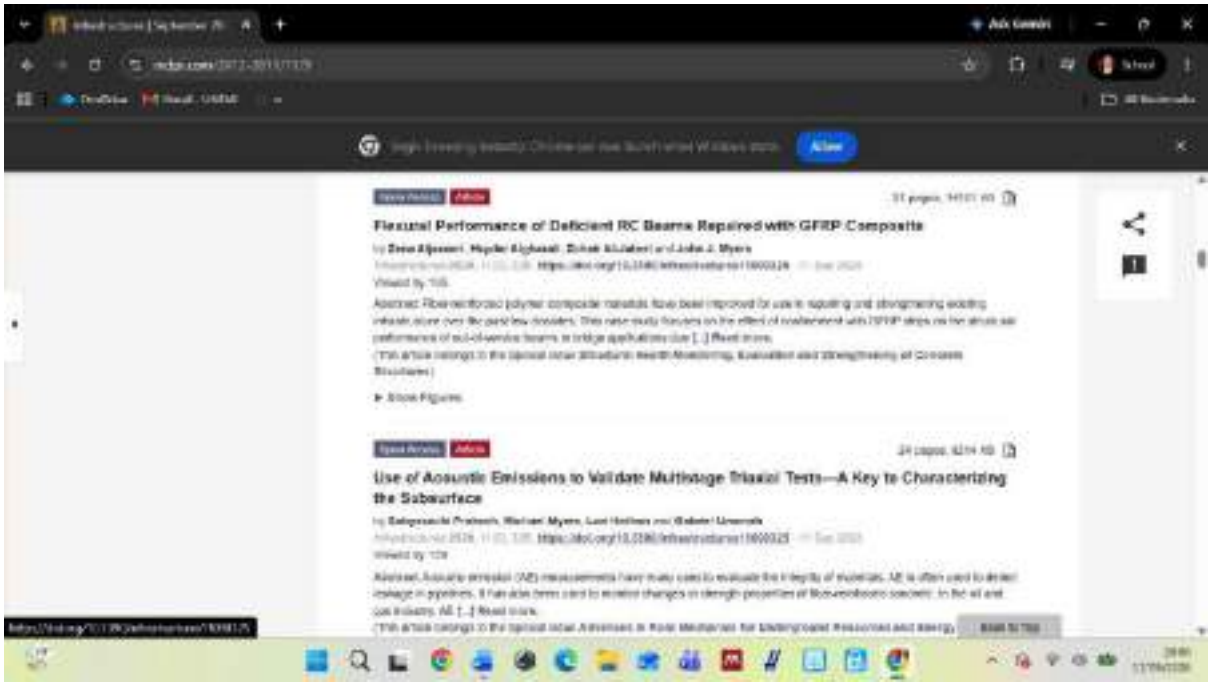
Vol 11 Issue 9 September 2026

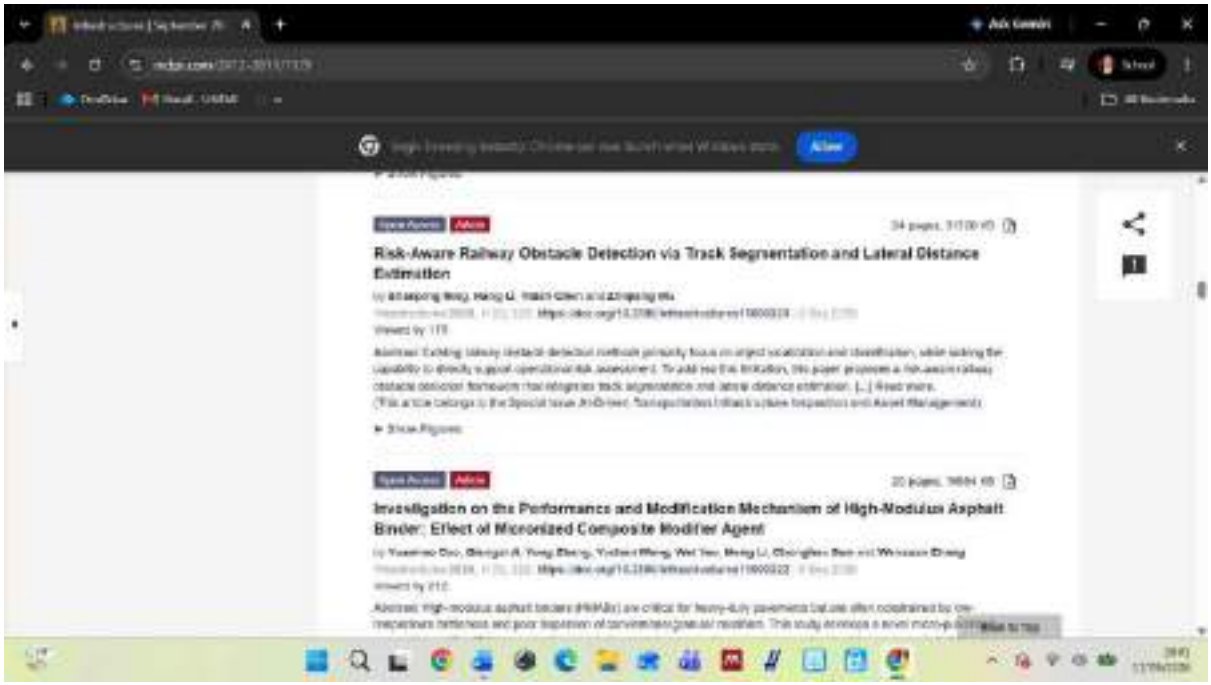


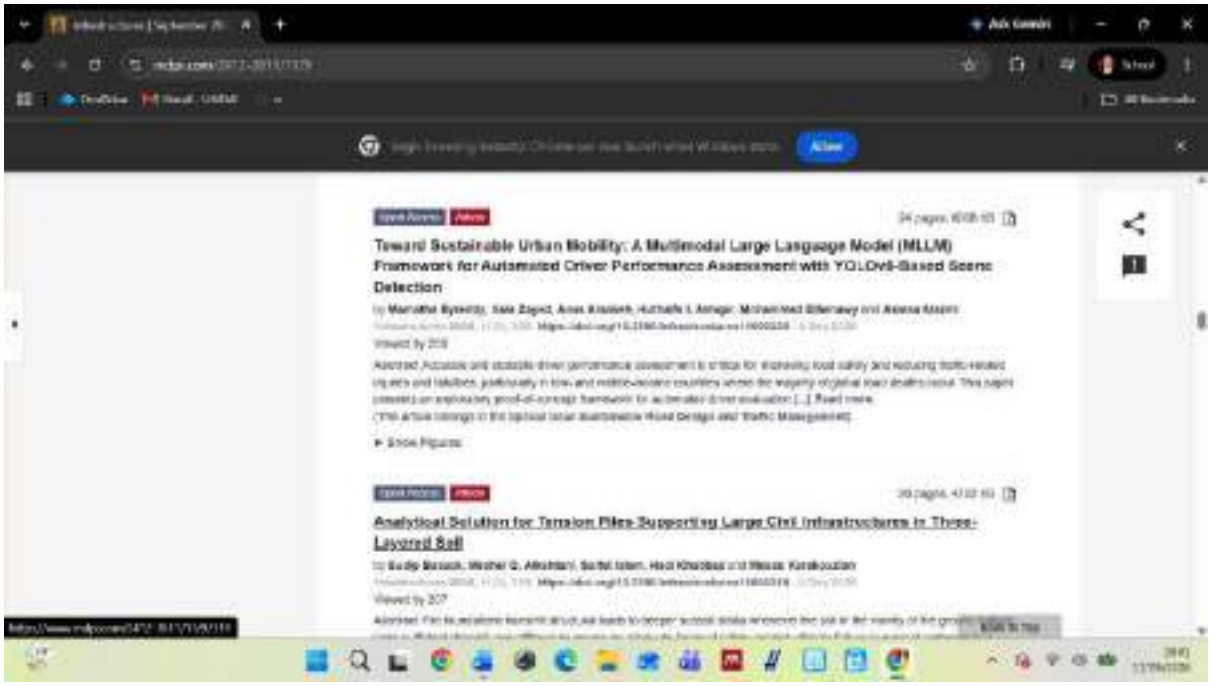


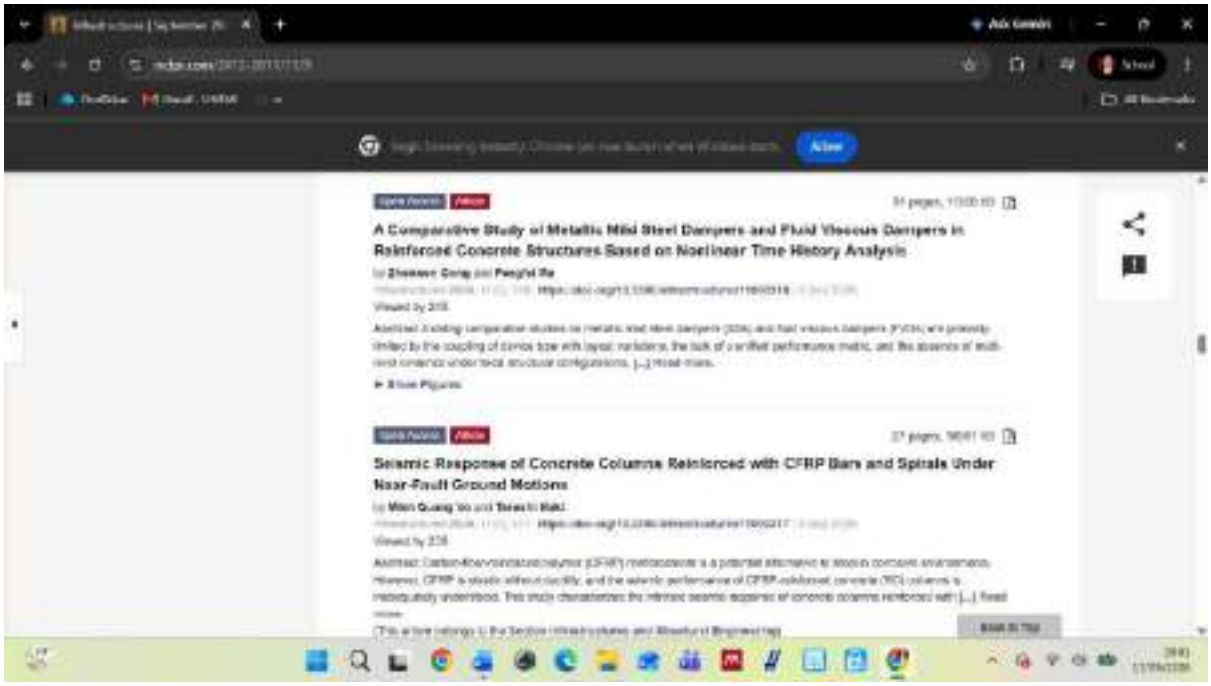

Journal Browser
River Corridors and River Engineering Works as Critical Infrastructure for Flood Resilience: Case Study of the Gradaščica River in Ljubljana, Slovenia
Pre-Event-Reference-Free TLS-Based Geometric Condition Assessment of Post-Fire RC Flexural Members

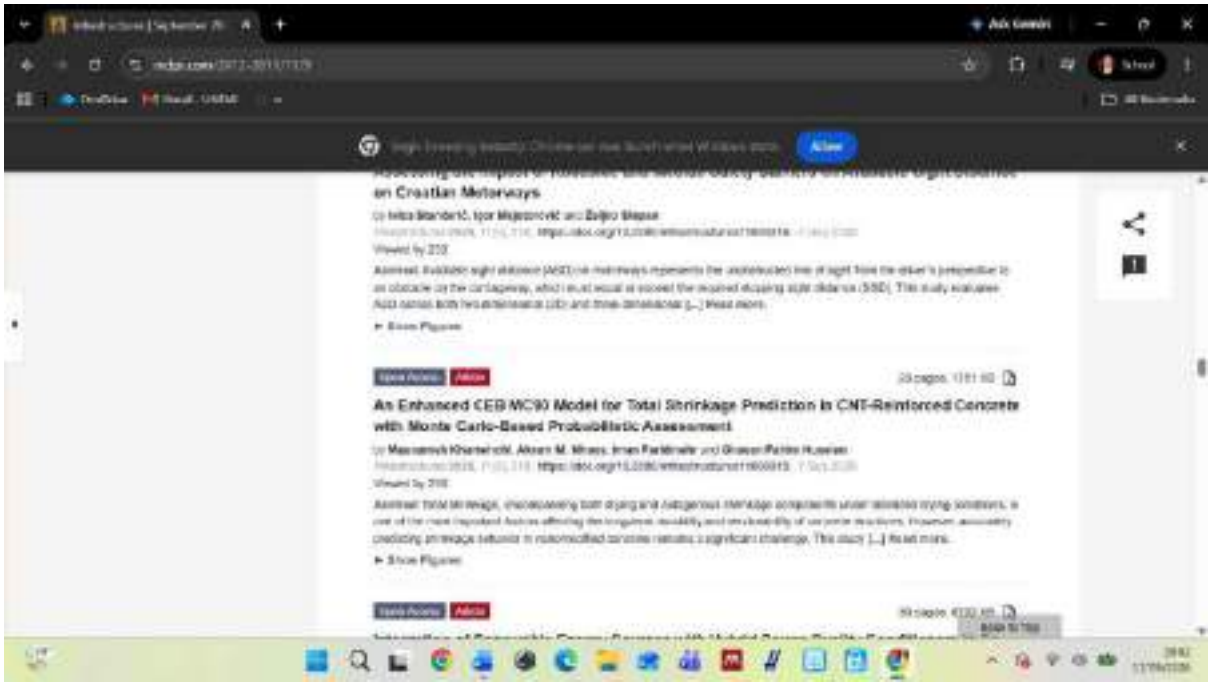

Affiliated Society:
Application of Gradient Boosting for Predicting Cost Overrun in Construction Projects
An Evaluation Framework for Social Return on Investment (SROI) of Low-Cost Road Safety Engineering Interventions

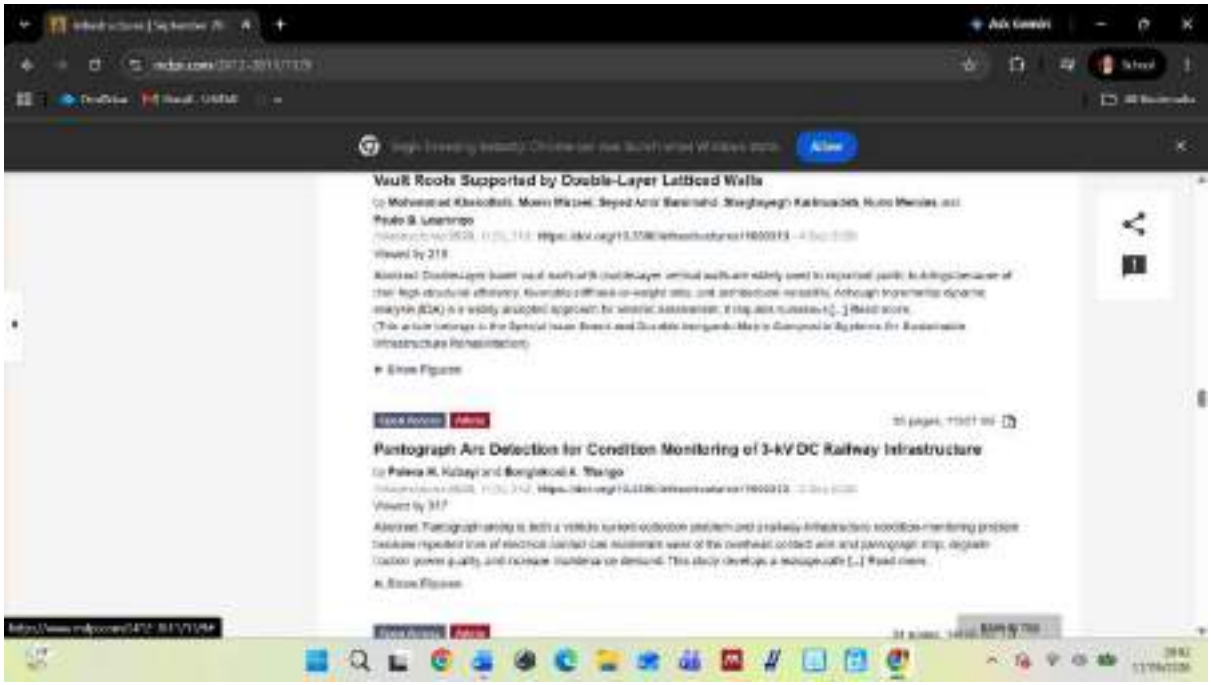

Flexural Performance of Deficient RC Beams Repaired with GFRP Composite
Use of Acoustic Emissions to Validate Multistage Triaxial Tests—A Key to Characterizing the Subsurface

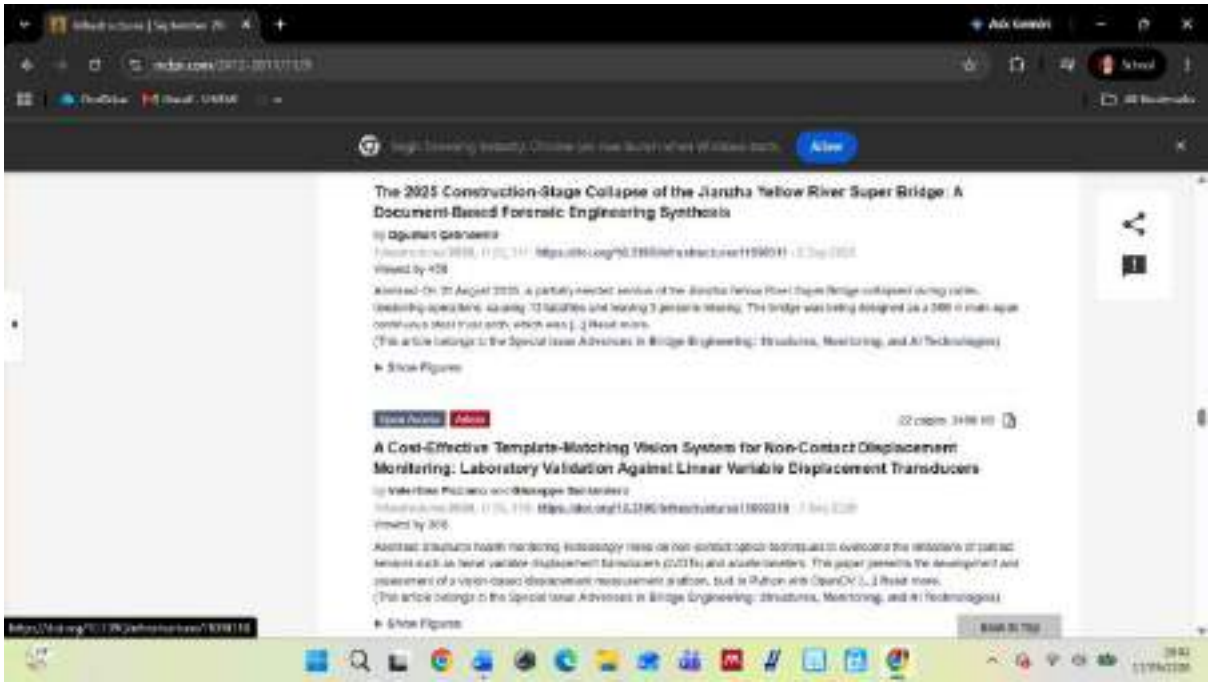

Risk-Aware Railway Obstacle Detection via Track Segmentation and Lateral Distance Estimation
Investigation on the Performance and Modification Mechanism of High-Modulus Asphalt Binder: Effect of Micronized Composite Modifier Agent

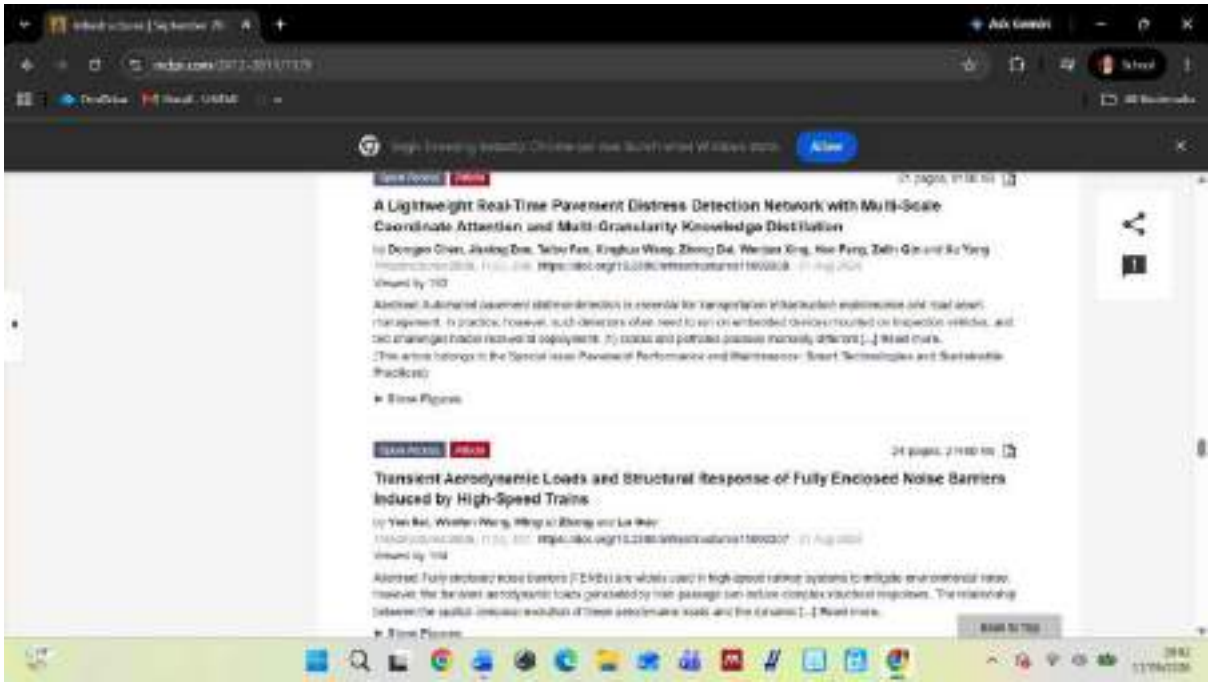

Toward Sustainable Urban Mobility: A Multimodal Large Language Model (MLLM) Framework for Automated Driver Performance Assessment with YOLOv8-Based Scene Detection
Analytical Solution for Tension Piles Supporting Large Civil Infrastructures in Three-Layered Soil

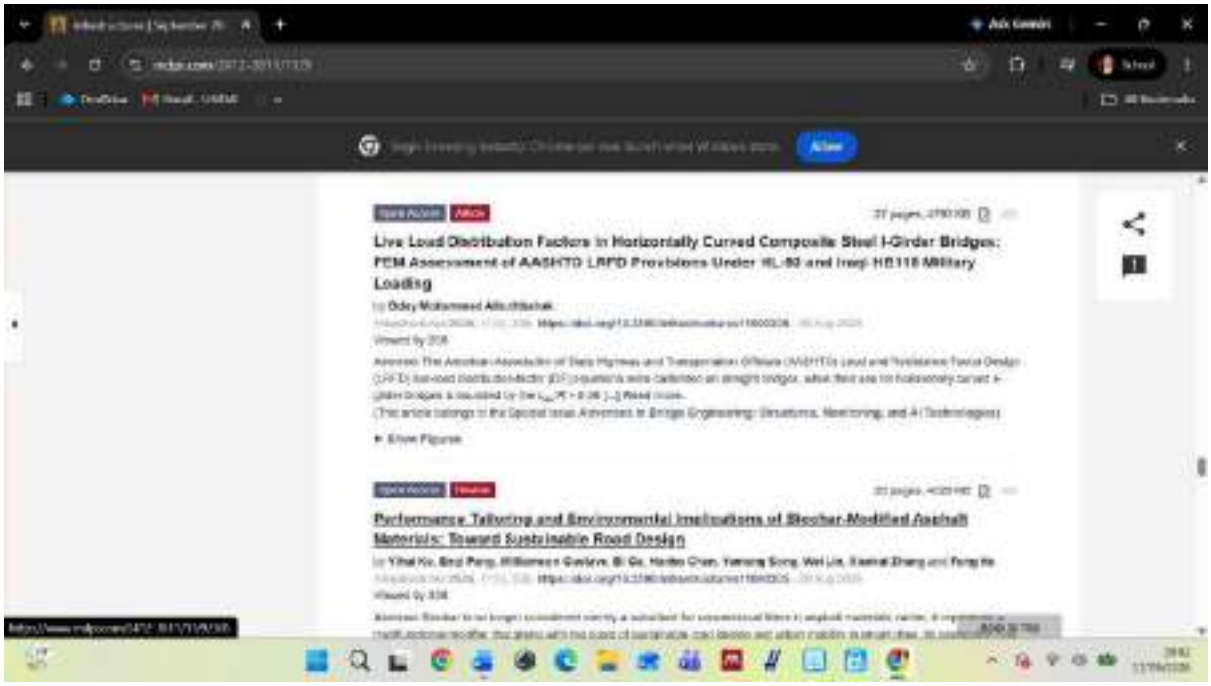

A Comparative Study of Metallic Mild Steel Dampers and Fluid Viscous Dampers in Reinforced Concrete Structures Based on Nonlinear Time History Analysis
Seismic Response of Concrete Columns Reinforced with CFRP Bars and Spirals Under Near-Fault Ground Motions

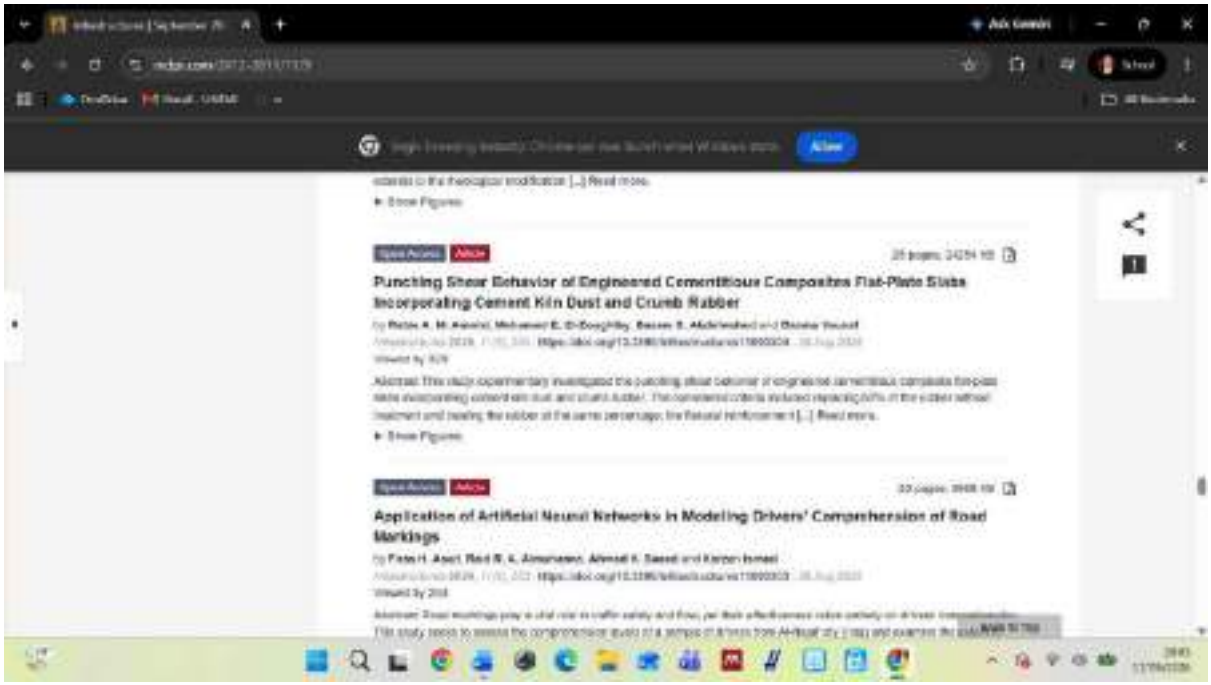

on Croatian Motorways
An Enhanced CEB MC90 Model for Total Shrinkage Prediction in CNT-Reinforced Concrete with Monte Carlo-Based Probabilistic Assessment

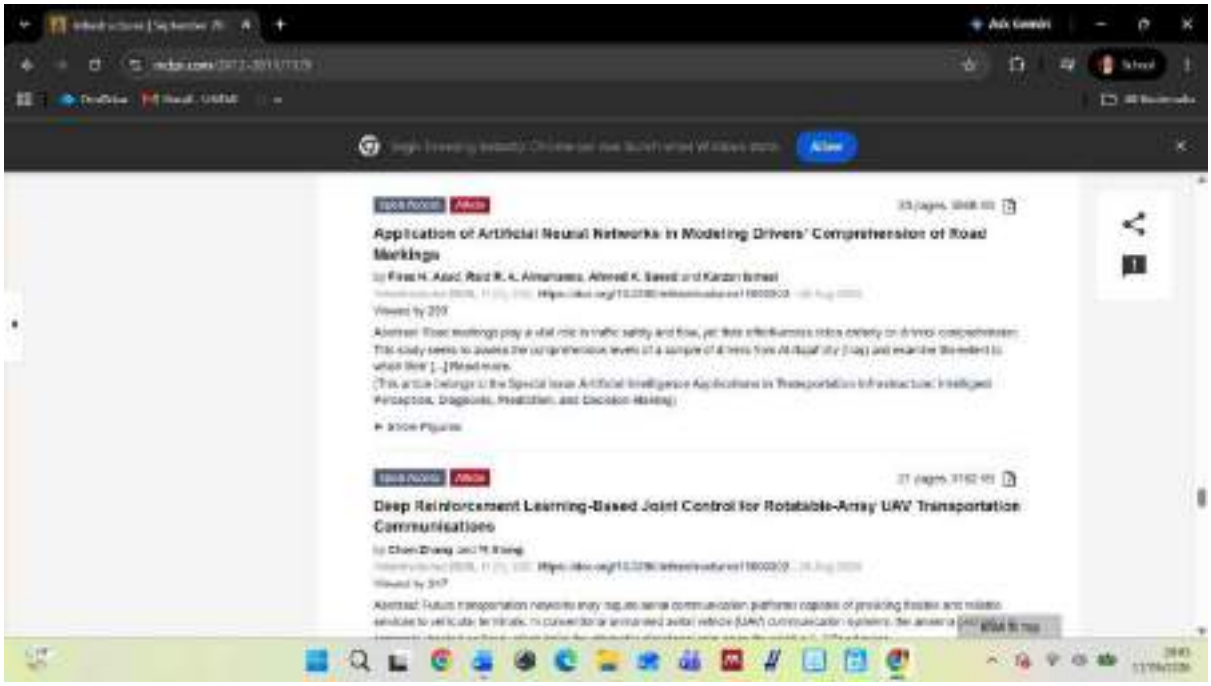

Vault Roofs Supported by Double-Layer Latticed Walls
Pantograph Arc Detection for Condition Monitoring of 3-kV DC Railway Infrastructure

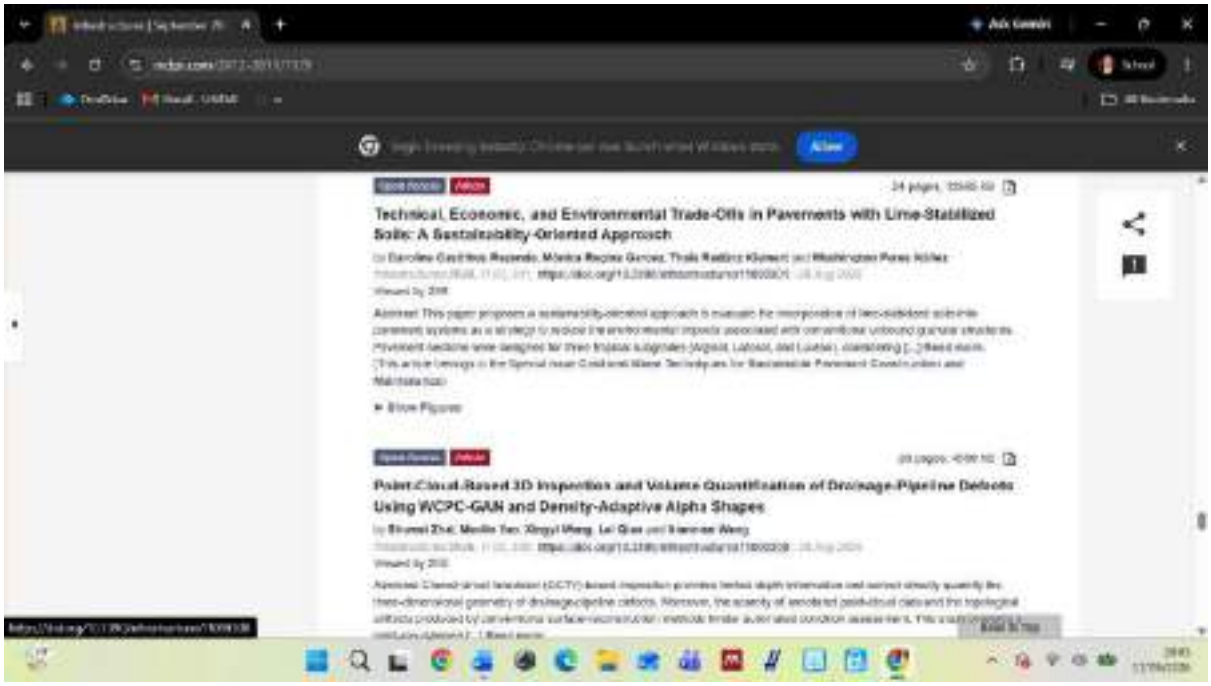

The 2025 Construction-Stage Collapse of the Jiansha Yellow River Super Bridge: A Document-Based Forensic Engineering Synthesis
A Cost-Effective Template-Matching Vision System for Non-Contact Displacement Monitoring: Laboratory Validation Against Linear Variable Displacement Transducers

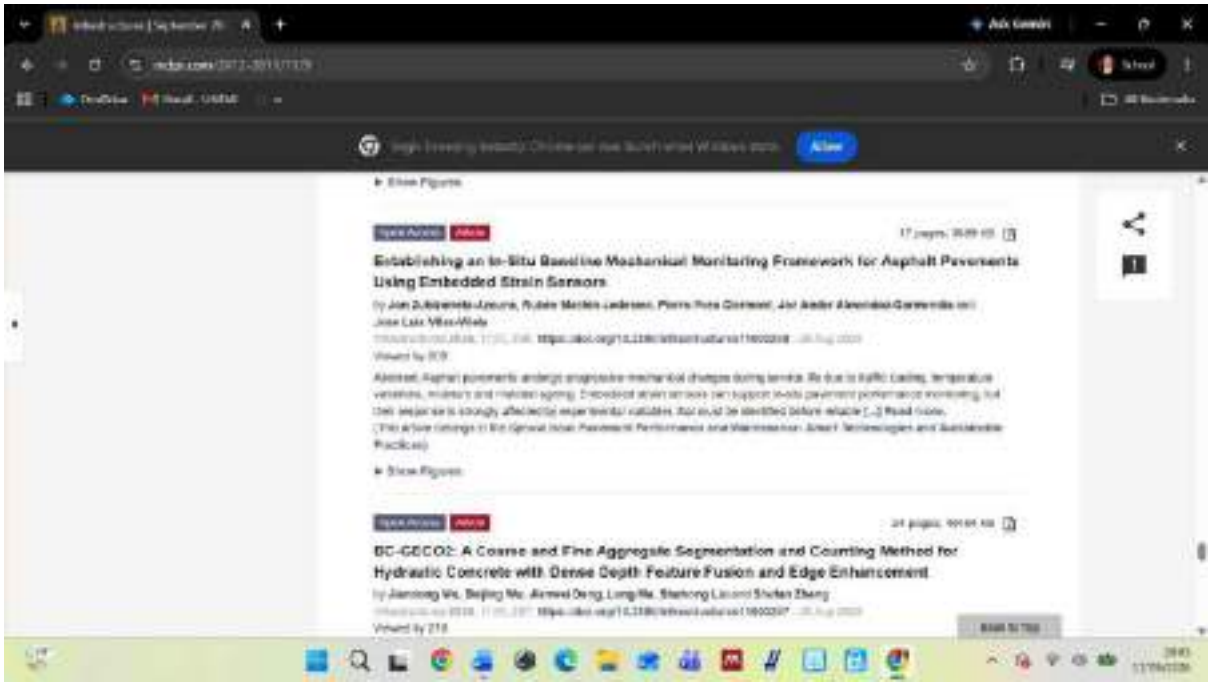

A Lightweight Real-Time Pavement Distress Detection Network with Multi-Scale Coordinate Attention and Multi-Granularity Knowledge Distillation
Transient Aerodynamic Loads and Structural Response of Fully Enclosed Noise Barriers Induced by High-Speed Trains

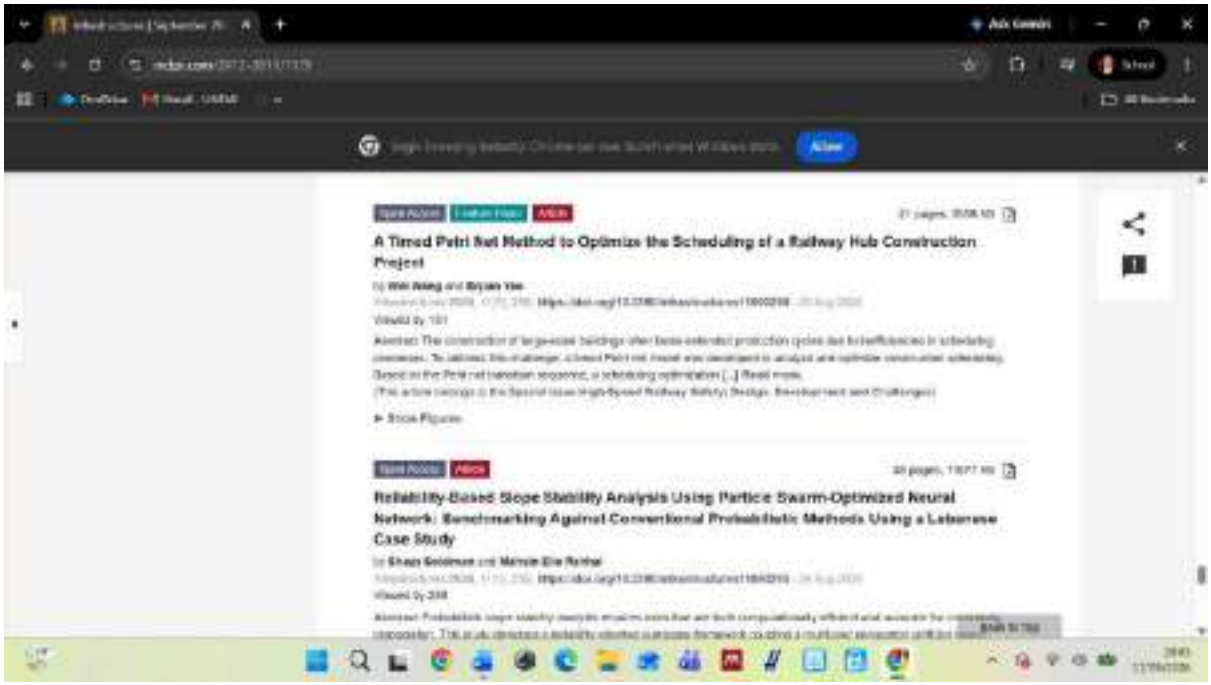

Live Load Distribution Factors in Horizontally Curved Composite Steel I-Girder Bridges: FEM Assessment of AASHTO LRFD Provisions Under HL-93 and Iraqi HB110 Military Loading
Performance Tailoring and Environmental Implications of Biochar-Modified Asphalt Materials: Toward Sustainable Road Design

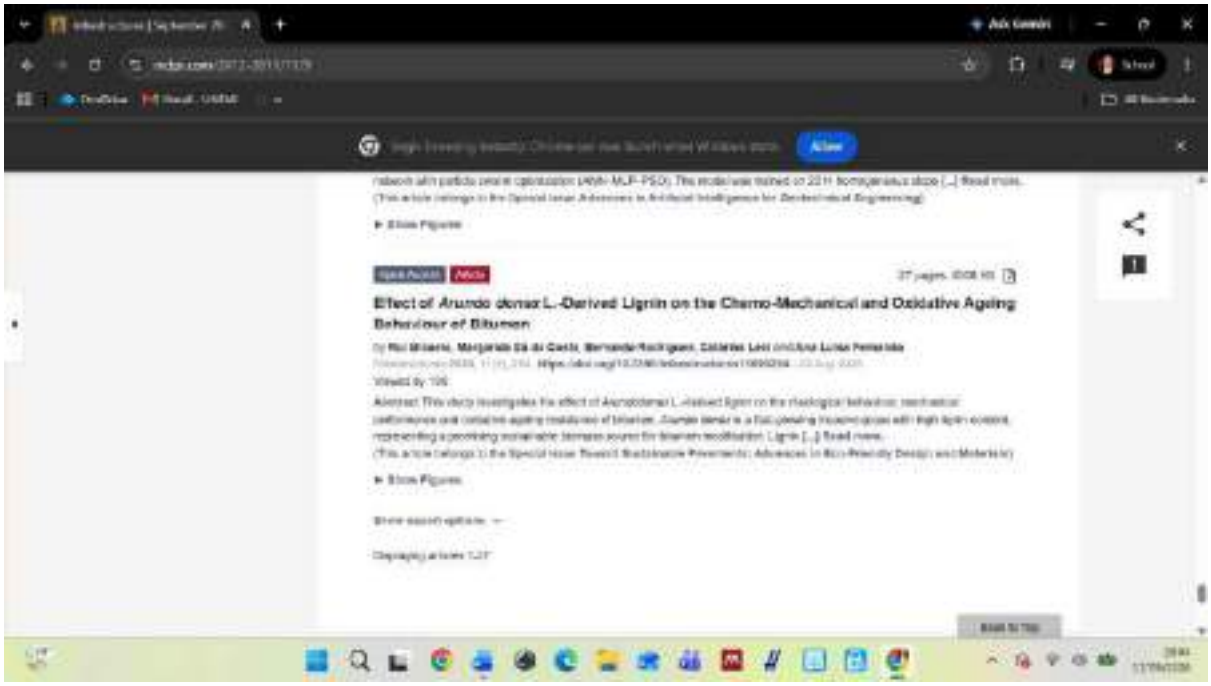

Punching Shear Behavior of Engineered Cementitious Composites Flat-Plate Slabs Incorporating Cement Kiln Dust and Crumb Rubber
Application of Artificial Neural Networks in Modeling Drivers' Comprehension of Road Markings

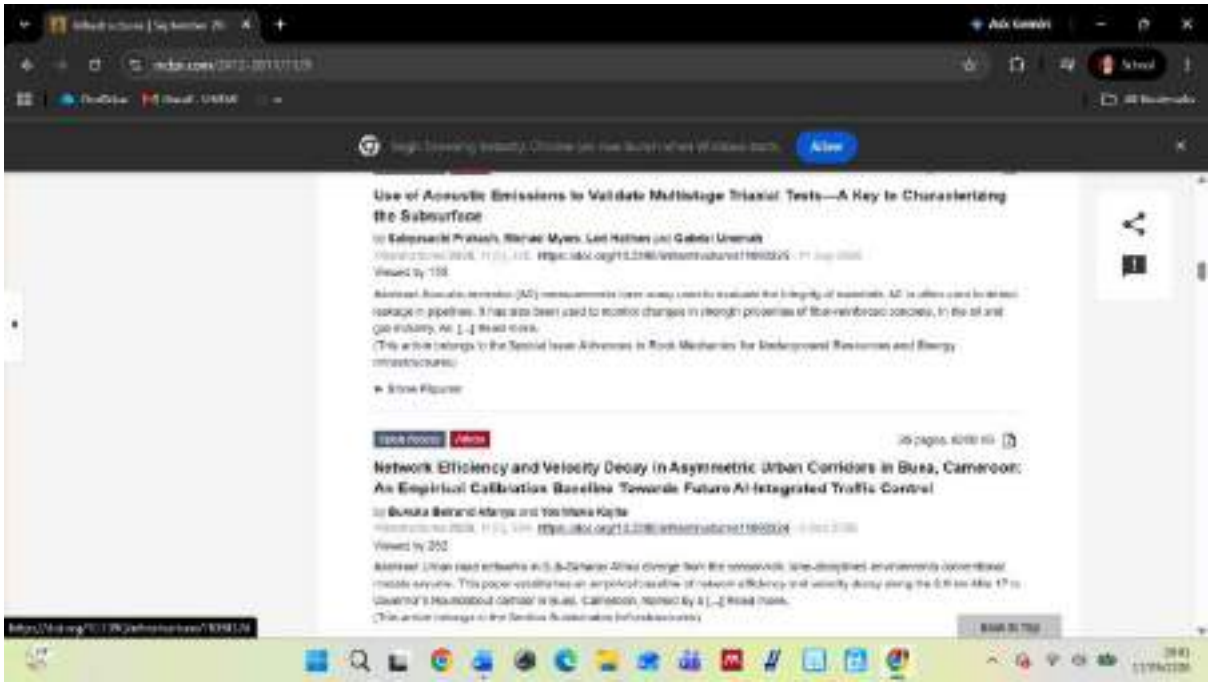

Application of Artificial Neural Networks in Modeling Drivers' Comprehension of Road Markings
Deep Reinforcement Learning-Based Joint Control for Rotatable-Array UAV Transportation Communications


Technical, Economic, and Environmental Trade-Offs in Pavements with Lime-Stabilized Soils: A Sustainability-Oriented Approach
Point-Cloud-Based 3D Inspection and Volume Quantification of Drainage-Pipeline Defects Using WCPC-GAN and Density-Adaptive Alpha Shapes


Establishing an In-Situ Baseline Mechanical Monitoring Framework for Asphalt Pavements Using Embedded Strain Sensors
BC-GECO2: A Coarse and Fine Aggregate Segmentation and Counting Method for Hydraulic Concrete with Dense Depth Feature Fusion and Edge Enhancement


A Timed Petri Net Method to Optimize the Scheduling of a Railway Hub Construction Project
Reliability-Based Slope Stability Analysis Using Particle Swarm-Optimized Neural Network: Benchmarking Against Conventional Probabilistic Methods Using a Lebanese Case Study


Effect of Arundo donax L.-Derived Lignin on the Chemo-Mechanical and Oxidative Ageing Behaviour of Bitumen


Use of Acoustic Emissions to Validate Multistage Triaxial Tests—A Key to Characterizing the Subsurface
Network Efficiency and Velocity Decay in Asymmetric Urban Corridors in Buea, Cameroon: An Empirical Calibration Baseline Towards Future AI-Integrated Traffic Control