INTERNATIONAL JOURNAL OF SUSTAINABLE CONSTRUCTION ENGINEERING AND TECHNOLOGY
ISSN: 2180-3242 e-ISSN: 2600-7959

IJSCET

Vol. 16 No. 1 (2025) 133-144
https://publisher.uthm.edu.my/ojs/index.php/ijscet

Evaluating Machine Learning Models for Predicting Structural Subsidence Based on Leveling Data: A Case Study in Vietnam

Nguyen Dinh Huy[1], Luong Ngoc Dung[1]*, Tran Dinh Trong[1], Tran Thi Hue[1]

[1] *Department of Geodesy and Geomatics,*
Hanoi University of Civil Engineering, Hanoi, 100000, VIETNAM

*Corresponding Author: dungln@huce.edu.vn
DOI: https://doi.org/10.30880/ijscet.2025.16.01.011

Article Info

Received: 27 March 2025
Accepted: 19 May 2025
Available online: 30 June 2025

Keywords

Machine learning models, performance evaluation, prediction, construction subsidence, leveling data, Vietnam

Abstract

Monitoring and predicting structural subsidence is crucial for construction project safety and efficiency, particularly in regions like Vietnam where leveling-based monitoring is standard. This study evaluates four machine learning methods, Linear Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB) for predicting subsidence using leveling data from a high-rise building in Hanoi. The dataset comprised 11 measurement cycles from three monitoring points between July 2020 and September 2022. Performance metrics revealed RF as the most effective model, consistently yielding superior predictive accuracy. LR also demonstrated steady, practical performance. Conversely, GB and SVM performed poorly, likely constrained by the limited dataset size. Notably, RF showcased the potential to surpass traditional predictive approaches, offering a robust solution even with sparse data, while LR remains a viable option for resource-constrained scenarios. This research introduces a modern ML-based approach to subsidence prediction relevant to the Vietnamese context, highlighting the importance of dataset characteristics and underscoring the need for larger datasets and the inclusion of more influencing factors in future investigations to further refine predictive capabilities.

1. Introduction

Structural subsidence is a critical factor that requires continuous monitoring throughout the lifecycle of civil and industrial constructions, from the construction phase to operational use. Subsidence refers to the vertical displacement of a foundation or the entire structure due to loading, geological changes, or environmental influences [1]. In civil engineering, structural subsidence is typically categorized into two main types: immediate subsidence (occurring instantly upon load application) and long-term subsidence (developing over time due to soil consolidation). Notably, during the construction phase, immediate subsidence often manifests as loads are applied to the ground, whereas long-term subsidence emerges gradually during operation, particularly in areas with weak soil foundations, such as clay or mud. For large-scale structures, such as high-rise buildings, bridges, hydroelectric dams, or tunnels, uncontrolled subsidence can lead to severe issues, including structural cracking, foundation instability, or even complete collapse. Therefore, accurately predicting subsidence mitigates risks and facilitates effective maintenance and repair planning.

To monitor and predict subsidence, accurately measuring subsidence is a fundamental prerequisite. Among the available measurement techniques, geometric leveling, commonly called the leveling method, is regarded as a traditional yet widely used approach due to its high precision [1]–[3]. This method employs a leveling instrument

This is an open access article under the CC BY-NC-SA 4.0 license.

and a rod to periodically measure elevation differences between subsidence monitoring points attached to the structure and reference benchmarks located outside the structure, thereby determining the elevation of the monitoring points. The elevation difference between two consecutive measurement cycles at a given monitoring point represents the subsidence of that point, enabling the calculation of the structure's subsidence parameters [2]. Technological advances have further enhanced the popularity of leveling, particularly with the development of electronic leveling instruments paired with barcode rods and integrated measurement software. Globally, this technique is widely applied to monitor subsidence in large-scale structures such as sea-crossing bridges, high-rise buildings, and infrastructure projects requiring high accuracy. In Vietnam, according to legal regulations and construction standards, subsidence measurement using the leveling method is a mandatory requirement for large-scale civil and industrial structures prone to subsidence, spanning from the construction phase through to the completion and operational phases as per Vietnam National Construction Standard [4]–[6]. Measurement cycles are typically conducted periodically (e.g., monthly, quarterly, or annually), depending on technical requirements and the construction timeline, to promptly detect anomalous changes and provide early warnings.

In Vietnam, predicting subsidence based on leveling measurement data commonly involves mathematical functions such as polynomial, exponential, hyperbolic, or Asaoka functions [7], [8]. These mathematical models assume that subsidence varies over time according to a specific pattern, enabling subsidence calculation at any predicted time point. However, this approach often lacks flexibility when handling leveling data affected by noise or irregular measurement frequencies. Additionally, it struggles to provide accurate long-term predictions when a subsidence dataset is insufficiently large or when the structure is influenced by anomalous factors, such as changes in groundwater levels, earthquakes, or nearby construction activities [8].

The advancement of Machine Learning (ML) has introduced novel approaches to processing and predicting structural subsidence, aiming to address the inherent limitations of traditional mathematical models which often lack flexibility with noisy or irregularly sampled data. Recent international studies highlight the efficacy of various ML techniques; for instance, Liu and Macedo [9] demonstrated the superiority of Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR) over conventional methods for liquefaction-induced settlements. Similarly, Zhang et al. [10] and Khajehzadeh et al. [11] have reported strong performance from deep learning models like Long Short-Term Memory (LSTM) networks and ANNs for general ground subsidence and shallow foundation settlement, respectively, underscoring their capability in handling complex, non-linear time-series data. A critical aspect emerging from the literature is the significant performance enhancement achieved via hybrid models, where hyperparameter optimization, as shown by Khajehzadeh et al. [11], substantially improves the predictive accuracy of base ML models. While these applications, often utilizing extensive datasets from InSAR or numerical simulations, showcase ML's potential, their direct applicability to the often sparse and discrete leveling data prevalent in Vietnamese construction monitoring remains an area requiring further investigation. Therefore, the curent work, therefore, seeks to evaluate established ML models under such data-constrained conditions typical in the Vietnam practice.

Applying ML to subsidence prediction addresses the limitations of traditional methods [12], [13]**.** Since subsidence measurement data over time often comprises a limited dataset, ML models can leverage their robust and superior performance capabilities. Recent studies have demonstrated that ML techniques achieve higher accuracy in interpolating and predicting subsidence; however, most research has focused on large-scale surface subsidence using data from GNSS (Global Navigation Satellite System) or InSAR (Interferometric Synthetic Aperture Radar) [14]–[17]. In Vietnam, the application of ML for subsidence prediction remains limited. Notable examples include a study employing an Artificial Neural Network (ANN) to predict subsidence in a hydroelectric structure using leveling data, achieving interpolation results closely aligned with actual subsidence, though predictions deviated by 3.2 mm [18]. Similarly, Random Forest (RF) and Support Vector Machine (SVM) models have proven effective in predicting subsidence based on leveling data from a single monitoring point [8]. Nevertheless, studies applying ML to process leveling-based subsidence data remain scarce. This article investigates advanced ML models to interpolate and predict subsidence in a structure using periodic leveling measurements.

In this research, four ML models, Linear Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB), were selected to evaluate their ability to handle nonlinear and complex data. The input data consists of subsidence measurements from 11 monitoring cycles, derived from leveling observations at three selected points within the subsidence monitoring network of the Dormitory–Canteen building at the School of Training and Professional Development in Auditing, located in the Hoa Lac High-Tech Park, Hanoi, Vietnam. These data reflect subsidence variations from the construction phase through to operational use. The study compares the performance of these models to identify the optimal approach while assessing the feasibility of applying ML to predict subsidence using real-world data. Additionally, the research aims to analyze the impact of data characteristics, specifically the limited number of cycles (11) and monitoring points (3), on model performance. The results provide a basis for selecting the most suitable ML model for subsidence prediction and contribute to improving high-rise building monitoring processes based on leveling data, offering practical application prospects nationally and beyond.

Int. J. of Sustainable Construction Engineering and Tech. Vol. 16 No. 1 (2025) p. 133-144

2. Data and Research Method

2.1 Data

This study utilizes subsidence measurement data from a high-rise building, specifically the Dormitory–Canteen project of the School of Training and Professional Development in Auditing. The project is managed by the Specialized Construction Investment Project Management Board under the State Audit Office of Vietnam. The Dormitory–Canteen building is a 12-story reinforced concrete frame structure with a pile foundation system (Fig. 1). Located within the Education and Training Zone of the Hoa Lac High-Tech Park in Hanoi, Vietnam, the site benefits from fully developed technical infrastructure and is characterized by the tropical climate of northern Vietnam. Surrounding facilities, including lecture halls, a stadium, and office buildings, with the tallest reaching 21 stories, have been completed and are operational. The area features flat terrain, stable geological conditions, and no recorded seismic activity.

Fig. 1 *Location of the Dormitory–Canteen Building of the School of Training and Professional Development in Auditing*

Subsidence monitoring for the Dormitory–Canteen building commenced during the completion phase and continued through the operational phase. The technical monitoring plan, including the number of measurement cycles and their timing, was designated by the main construction contractor under Vietnam's National Technical Standard TCVN 9360 [4]. The first measurement cycle was conducted on July 10, 2020, and the final cycle, the 11th, was completed on September 29, 2022. The initial five cycles, measured during the completion phase, occurred at an average interval of one month, while the subsequent six cycles, taken during the operational phase, had significantly varying intervals. The subsidence monitoring network comprises 41 points, labeled M1, M2, ..., M41 (marked green in Fig. 2), positioned at the building's load-bearing structural elements (Fig. 2) with a density compliant with TCVN 9360 [4]. This network is linked to a reference benchmarks network of three points outside the Dormitory–Canteen project area. The subsidence monitoring network was measured with second-order leveling accuracy [4].

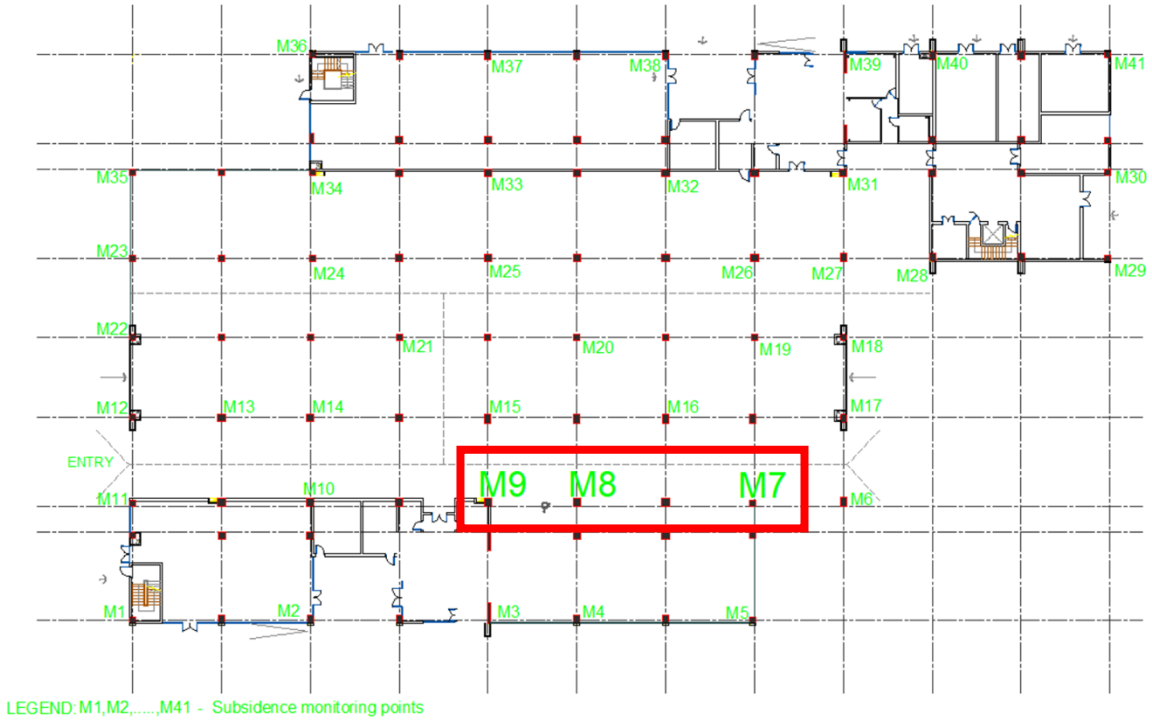

Fig. 2 *Location diagram of subsidence monitoring points and three points M7, M8, and M9 selected for the experiment*

In this study, subsidence data from three points M7, M8, and M9 (highlighted in bold red square in Figure 2) were selected for analysis due to their proximity to each other and continuous measurement throughout the monitoring period, unlike other points that experienced interruptions. The raw subsidence values (in mm) recorded for these three points across the 11 measurement cycles, along with their corresponding measurement dates, are presented in Table 1. This table clearly illustrates the progressive increase in settlement observed at each monitoring location over time.

Table 1 *Subsidence data for points M7, M8, and M9 across 11 cycles*

Cycle	Time	Subsidence (mm)			Cycle	Time	Subsidence (mm)		
		M7	M8	M9			M7	M8	M9
1st	2020-07-10	0	0	0	7th	2021-03-30	-3.64	-7.21	-8.32
2nd	2020-08-12	-0.91	-1.24	-1.36	8th	2021-06-28	-4.28	-7.73	-8.87
3rd	2020-09-01	-1.37	-2.51	-2.99	9th	2021-09-29	-5.14	-8.68	-9.85
4th	2020-09-25	-1.94	-3.86	-4.06	10th	2022-06-26	-6.07	-9.56	-10.77
5th	2020-10-28	-2.47	-5.35	-6.09	11th	2022-09-29	-6.44	-9.87	-11.11
6th	2020-12-30	-3.62	-7.11	-8.14					

Table 1 highlights the cumulative settlement (in mm) at each point for each measurement cycle, demonstrating the overall subsidence trend from July 2020 to September 2022.

The data in Table 1 reveal a gradual increase in subsidence over time at all three points, with the largest subsidence recorded at M9 (-11.11 mm in cycle 11) and the smallest at M7 (-6.44 mm in cycle 11). The time intervals between cycles vary irregularly, ranging from approximately 20 days (between cycles 3 and 4) to over 9 months (between cycles 9 and 10), reflecting the practical realities of structural monitoring, where measurement frequency adjusts to the technical requirements specified by the main construction contractor.

2.2 Machine Learning Models

This study employs four ML models LR, SVM, RF, and GB to interpolate and predict subsidence based on leveling data from the Dormitory–Canteen high-rise building at the School of Training and Professional Development in Auditing. These models were selected for their diverse approaches (linear and nonlinear), ability to handle small datasets, and proven efficacy in similar engineering applications. Below is a brief description of each method and the rationale for its application in this study.

Linear Regression (LR) [19]: LR is a statistical method that models the linear relationship between independent and dependent variables by minimizing the least squares loss function. It is well-suited for predicting continuous values, assuming a linear data distribution. In this study, LR assumes a linear relationship between time (days since the first cycle) and subsidence. Its simplicity and ease of implementation make it an appropriate baseline to assess whether a linear function can effectively subsidence trends.

Support Vector Machine (SVM) [20]: SVM is a supervised learning model that constructs an optimal hyperplane to separate classes in feature space (for classification) or estimate continuous values (for regression). It employs kernel functions to handle nonlinear data, optimizing the maximum margin between classes. This study uses SVM is applied as Support Vector Regression (SVR) with a nonlinear kernel (e.g., Radial Basis Function, RBF) to model the nonlinear relationship between time and subsidence. SVM's suitability for small datasets, such as the 11-cycle dataset, and its capacity to manage potential noise in leveling measurements make it a valuable choice.

Random Forest (RF) [21]: RF is an ensemble learning method that combines multiple decision trees trained on random subsets of data (via bootstrapping) and random feature selections. Predictions are aggregated through majority voting (for classification) or averaging (for regression), offering high accuracy and robustness against noise. RF excels at capturing nonlinear patterns and variable interactions while being less sensitive to noisy data. In this study, RF is expected to leverage the gradual subsidence trend over time, even with a limited sample size.

Gradient Boosting (GB) [22]: GB is a sequential ensemble technique where weak learners (typically decision trees) are trained iteratively to minimize the residual errors of prior models, using gradient descent to optimize the loss function. It is known for its high accuracy but is sensitive to hyperparameter tuning. This study chose GB for its adaptability to irregular data (e.g., varying intervals between cycles) and its effectiveness in predicting long-term subsidence trends.

These models were trained separately for each monitoring point (M7, M8, and M9) to compare their performance in predicting subsidence based on time.

2.3 Experimental Procedure

The experimental procedure was designed to train, optimize, and evaluate the performance of ML models using subsidence data from the Dormitory–Canteen building. The workflow, illustrated in Fig. 3, encompasses the following steps:

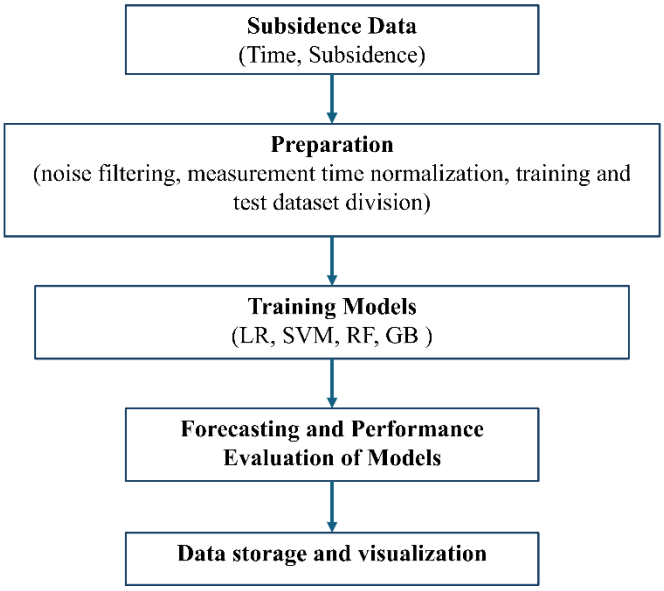

Fig. 3 *Process of calculating subsidence predictions using ML models*

Data preparation: The subsidence dataset was (1) inspected to ensure no missing values or outliers were present. Next, (2) measurement times were standardized as the number of days since the first cycle (July 10, 2020)

to create a continuous independent variable, resulting in cycle 2 at 33 days, cycle 3 at 53 days, and so forth. This standardization ensures a continuous independent variable and consistent scaling for the ML models. Finally, (3) the data were split into two sets: 73% (the first 8 cycles) for training the models and the remaining 27% (the last 3 cycles) for testing predictive performance. This division ensures sufficient data for model training while reserving a portion to assess generalization to unseen data points. The independent variable is the number of days since the first cycle, and the dependent variable is the subsidence recorded at points M7, M8, and M9.

Model training: Each model (LR, SVM, RF, GB) was trained independently on the training set using default configurations from Scikit-learn. This approach was adopted primarily due to the very limited dataset size (11 cycles, 3 points), which often makes extensive hyperparameter optimization prone to overfitting or yielding results that may not generalize well. While techniques like GridSearchCV could potentially refine model parameters, the risk of tailoring models too closely to the small training set was prioritized for avoidance in this initial investigation. Specifically, LR was implemented with LinearRegression(). SVM was deployed as Support Vector Regression (SVR) via SVR (kernel='rbf'), employing an RBF kernel with C=1.0 and gamma='scale'. RF was initialized with RandomForestRegressor(n_estimators=100, max_depth=None), using 100 trees and no depth limit. GB was configured with GradientBoostingRegressor(n_estimators=100, learning_rate=0.1), featuring 100 trees and a default learning rate of 0.1. Each model was trained separately for each monitoring point (M7, M8, M9) using the .fit() method, where X_train represents the ordinal days and y_train the corresponding subsidence values.

Prediction and performance evaluation: After training, the models predicted subsidence for the test set (the final 3 cycles) using the .predict() method, with X_val as the ordinal days of the test cycles. Performance was assessed using three metrics: Mean Squared Error (MSE) from mean_squared_error(), Mean Absolute Error (MAE) from mean_absolute_error(), and Coefficient of Determination (R^2) from r2_score(). These metrics were computed by comparing predicted values to actual subsidence at the cycles of September 29, 2021, June 26, 2022, and September 29, 2022. Additionally, the models predicted subsidence across all 11 cycles to illustrate trends, though only the last 3 were used for performance evaluation.

Result storage and visualization: Results for each point were saved to text files (e.g., M7_result.txt) in an output directory containing dataset size, performance metrics (MSE, MAE, R^2), and predicted values for the 3 test cycles. Aggregate performance across all three points was recorded in a model_performance.csv file. Visualizations were generated using Matplotlib, plotting actual data (as points) alongside prediction curves for the four models, saved as PNG files (e.g., M7_plot.png) at 600 DPI resolution. These plots visually compare each model's predictive capability across the full-time series.

This procedure was repeated independently for each point (M7, M8, M9) to evaluate model effectiveness across varying subsidence patterns. Experimental results, including performance metrics and predicted subsidence values, are detailed in Section 5. These findings provide a basis for comparing and analyzing the efficacy of LR, SVM, RF, and GB.

This workflow was implemented in the open-source Python programming language, utilizing libraries such as Scikit-learn for ML, NumPy and SciPy for numerical computations, and Matplotlib for data visualization. The complete code and data used are available at the DOI link 10.17605/OSF.IO/ZWP4A.

3. Results

The performance of these models was evaluated on the test dataset, comprising the last three measurement cycles. Figure 4 visually compares the actual subsidence values with the predictions made by each of the four ML models (LR, SVM, RF, and GB) for monitoring points M7, M8, and M9 over these test cycles. The plots allow for a direct assessment of how closely each model's predictions align with the observed field data.

Figure 4 shows the comparison of actual and predicted subsidence by ML models for points M7, M8, and M9 over the three test cycles (September 2021, June 2022, September 2022). Each panel illustrates the actual subsidence (points) alongside the predicted subsidence trend lines for LR, SVM, RF, and GB, allowing for visual evaluation of model fit.

The performance of four ML models (LR, SVM, RF, and GB) in predicting subsidence at points M7, M8, and M9, based on data from 11 leveling measurement cycles, was evaluated on the test dataset (comprising the last three cycles: September 29, 2021, June 26, 2022, and September 29, 2022) using three metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2). The results are presented in Table 2.

Int. J. of Sustainable Construction Engineering and Tech. Vol. 16 No. 1 (2025) p. 133-144

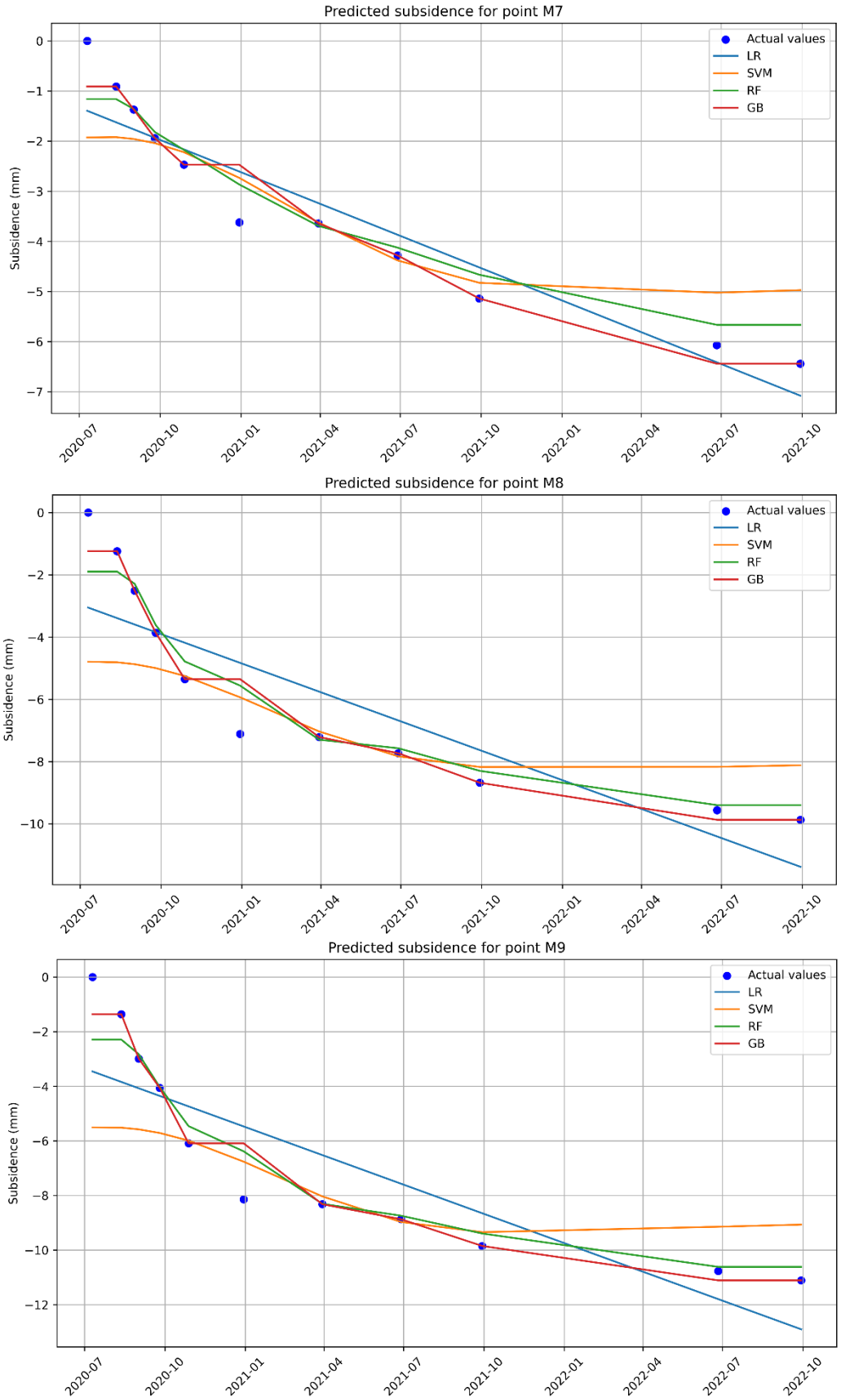

Fig. 4 Subsidence predicted by ML models of points M7, M8, and M9

Table 2 *Summary of performance evaluation of models*

Point	Metric	LR	SVM	RF	GB
M7	MSE	1.0289	5.3886	0.6986	2.7685
	MAE	0.9166	2.0750	0.7161	1.5165
	R^2	0.8345	0.1332	0.8876	0.5547
M8	MSE	5.0773	15.6339	2.0999	7.6132
	MAE	2.0592	3.3267	1.2460	2.5451
	R^2	0.6911	0.0490	0.8723	0.5369
M9	MSE	6.7074	20.3354	2.4954	9.7507
	MAE	2.3838	3.7650	1.3167	2.8905
	R^2	0.6809	0.0325	0.8813	0.5361

To further illustrate the comparative performance in terms of prediction error, Figure 5 presents a bar chart of the MAE for each ML model at the three monitoring points. This visualization clearly highlights the consistently lower MAE achieved by the RF model compared to LR, SVM, and GB, reinforcing its superior accuracy.

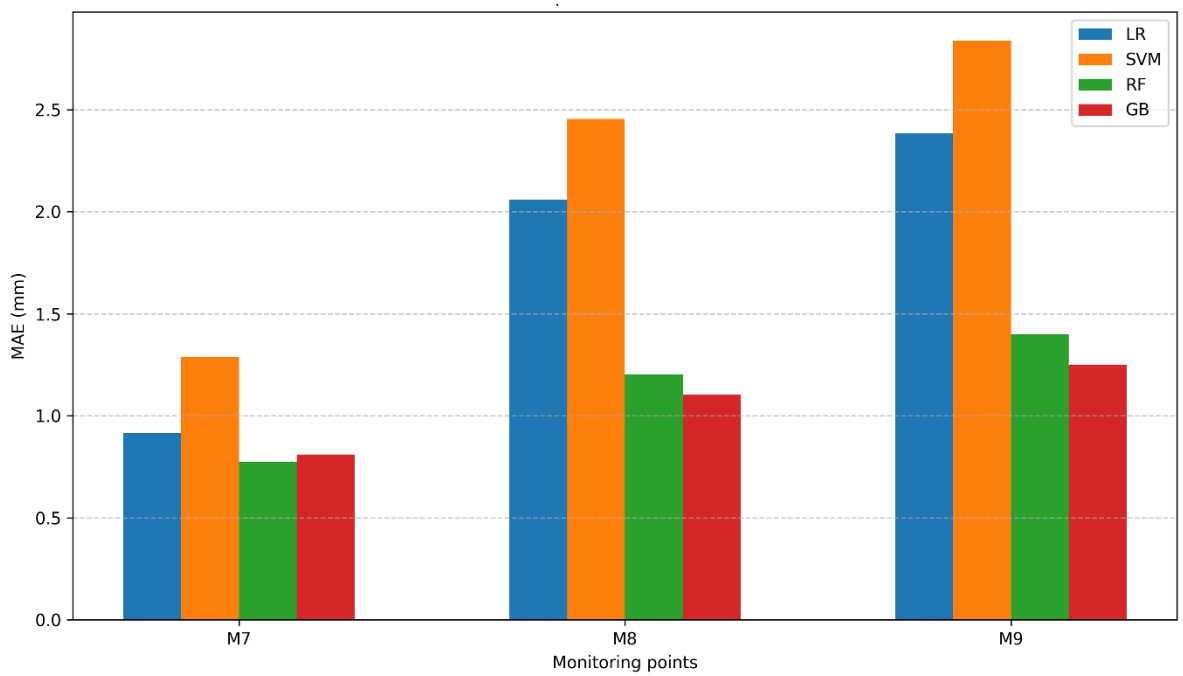

Fig. 5 *Comparison of MAE across ML models*

Figure 5 illustrates Comparison of Mean Absolute Error (MAE) across ML models for points M7, M8, and M9. The chart demonstrates that the Random Forest (RF) model consistently yields the lowest MAE, indicating higher predictive accuracy compared to Linear Regression (LR), Support Vector Machine (SVM), and Gradient Boosting (GB).

At point M7, RF achieved the best performance with an MSE of 0.6986, an MAE of 0.7161 mm, and an R^2 of 0.8876, indicating that the model explains nearly 89% of the variance in the real subsidence data. LR ranked second with an MSE of 1.0289, an MAE of 0.9166 mm, and an R^2 of 0.8345, demonstrating strong predictive capability despite being a simple linear model. GB exhibited moderate performance with an MSE of 2.7685, an MAE of 1.5165 mm, and an R^2 of 0.5547, while SVM performed the poorest with an MSE of 5.3886, an MAE of 2.0750 mm, and an R^2 of 0.1332, accounting for only approximately 13% of the data variance.

At point M8, RF continued to lead with an MSE of 2.0999, an MAE of 1.2460 mm, and an R^2 of 0.8723, confirming its superior predictive capability. LR exhibited acceptable performance with an MSE of 5.0773, an MAE of 2.0592 mm, and an R^2 of 0.6911, while GB performed less effectively with an MSE of 7.6132, an MAE of 2.5451 mm, and an R^2 of 0.5369. SVM again yielded the lowest performance with an MSE of 15.6339, an MAE of 3.3267 mm, and an R^2 of 0.0490, rendering it largely unsuitable for the data

At point M9, similarly, RF achieved the best results with an MSE of 2.4954, an MAE of 1.3167 mm, and an R^2 of 0.8813, further demonstrating its effectiveness on nonlinear data. LR recorded an MSE of 6.7074, an MAE of

Int. J. of Sustainable Construction Engineering and Tech. Vol. 16 No. 1 (2025) p. 133-144

2.3838 mm, and an R^2 of 0.6809, maintaining stable performance. GB exhibited an MSE of 9.7507, an MAE of 2.8905 mm, and an R^2 of 0.5361, while SVM delivered the poorest performance with an MSE of 20.3354, an MAE of 3.7650 mm and an R^2 of 0.0325, highlighting its significant limitations.

To provide a comparative visual representation of how well each model explains the variance in the observed subsidence, Figure 6 displays the Coefficient of Determination (R^2) values for each ML model across the three monitoring points (M7, M8, and M9). This chart allows for a quick assessment of model fit, with higher R^2 values closer to 1.0 indicating a better correspondence between predicted and actual subsidence.

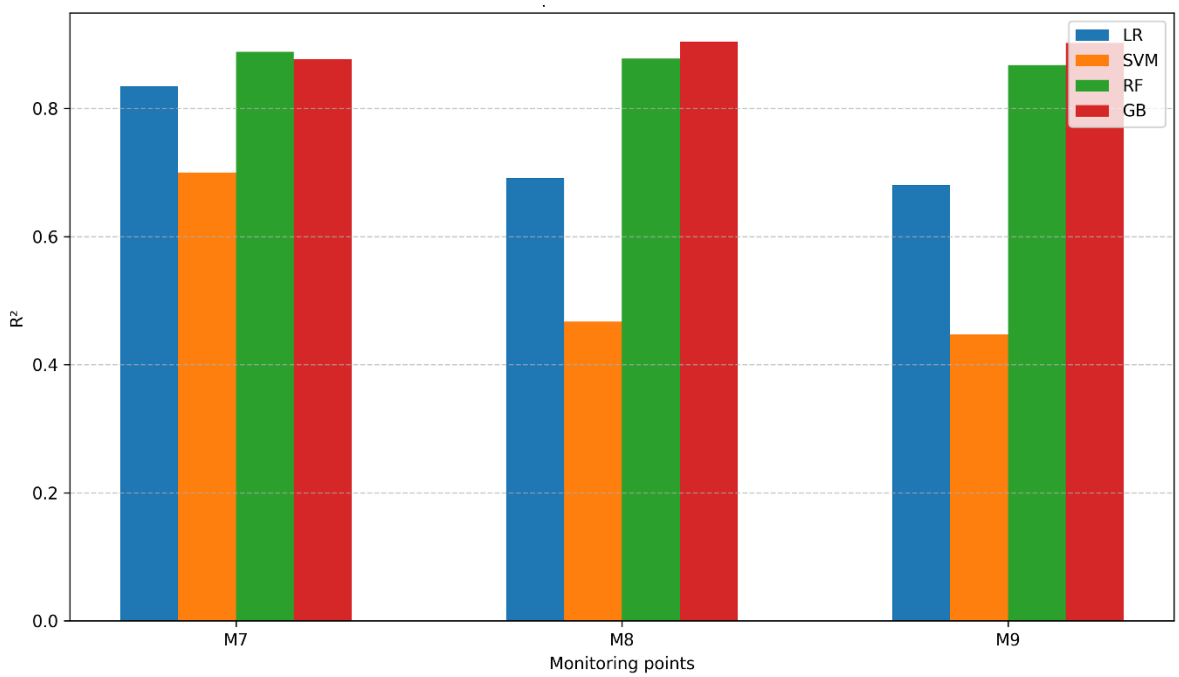

Fig. 6 *Comparison of R2 across ML models*

The Figure 6 illustrates that the Random Forest (RF) model consistently achieves the highest R^2 values, signifying a superior ability to explain the variance in the observed subsidence data compared to Linear Regression (LR), Support Vector Machine (SVM), and Gradient Boosting (GB).

Thus, the performance of the models reveals distinct differences in their ability to predict subsidence at points M7, M8, and M9, while also elucidating the influence of data characteristics on the outcomes. RF is the most optimal model across all three points, with an average MAE ranging from 1.0 to 1.3 mm and an R^2 fluctuating between 0.87 and 0.89. This can be attributed to RF's capability to capture the nonlinear trends of subsidence, which increased rapidly in the initial cycles (2020) and slowed in the later stages (2021–2022). Additionally, RF demonstrates lower sensitivity to the small size of the training dataset (8 measurement cycles of training dataset) due to its ensemble mechanism of multiple decision trees, which helps mitigate overfitting. The predicted subsidence results for the three points are presented in Table 3.

Table 3 *Predicted subsidence of the models compared to actual values*

Time	Actual value (mm)	LR (mm)	SVM (mm)	RF (mm)	GB (mm)
2021-09-29	-5.14	-4.5198	-3.3856	-4.6402	-4.0109
2022-06-26	-6.07	-6.4125	-3.3032	-5.9604	-4.5244
2022-09-29	-6.44	-7.0784	-3.2865	-5.9604	-4.5244

Table 3 illustrates the deviation between the predicted and actual subsidence values in the last three measurement cycles, demonstrating that RF performs best in cycles with longer time intervals (e.g., from September 29, 2021, to June 26, 2022). In contrast, LR and GB face challenges in long-term predictions (cycles 10 and 11), while SVM is unsuitable in all cases. These results confirm that RF is the optimal choice, whereas LR may be considered if computational simplicity and limited hardware resources are prioritized.

4. Discussion

This study focuses on comparing the ML models of LR, SVM, RF, and GB for predicting subsidence based on leveling measurement data from the monitoring network of the Dormitory–Canteen building at the School of Training and Professional Development in Auditing (Hanoi, Vietnam). The findings hold practical significance in subsidence monitoring in Vietnam, where periodic leveling measurements are mandatory.

The performance evaluation demonstrates that the RF model, with its superior performance (low MAE and high R^2), can be integrated into automated monitoring systems to predict subsidence using real-world leveling data. This is particularly valuable for super high-rise buildings, large bridges, or hydroelectric dams, where continuous subsidence monitoring is essential to ensure structural safety and facilitate timely maintenance planning.

In terms of computational resources, while formal benchmarking was not a primary objective due to the small dataset (11 cycles, 3 points) resulting in rapid training for all models, this aspect is crucial for larger-scale applications. Here, all models trained quickly, but for more extensive datasets, the known efficiency of RF and LR would be advantageous.

The practical significance also provides a flexible approach to analyzing the subsidence from leveling data with irregular frequencies, a common characteristic in practice. The RF model can be adapted to predict subsidence in long-term scenarios, aiding managers in making data-driven decisions rather than relying solely on design estimates.

The robustness of the RF model to the specific conditions of this study—minimal and irregularly spaced data is a key finding. It demonstrated a lower sensitivity to these data limitations compared to GB and SVM. A formal sensitivity analysis, exploring how model predictions change with systematic variations in input data (e.g., simulated noise levels or different degrees of data regularity), was not conducted but represents an important avenue for future research. Such an analysis would provide deeper insights into the operational envelope within which each model maintains reliable performance and further quantify their respective robustness.

Despite these promising results, the study acknowledges several limitations:

First, our dataset was quite small (11 measurement cycles from only 3 points). This is common for subsidence monitoring in our local area, but it definitely made it harder for our machine learning models to learn well and apply to other situations. This was especially true for the GB and SVM models, which usually need much more data to work best. This shortage of data is likely why GB didn't perform as well as we hoped (R^2 around 0.53–0.55) and why SVM couldn't really find the subsidence trends (R^2 below 0.15). Of course, getting more real measurements over a longer time would be best, but that's often difficult to do. So, for future projects with small datasets like this, we could try a few things. For example, we could use techniques to create more training data from the data we already have. This is called synthetic data generation or data augmentation, using methods like bootstrapping. Another idea is to use good time-series forecasting methods to help simulate what longer data might look like, which could help train the models better at the start. If we had a larger or artificially expanded dataset, these GB and SVM models would likely perform much better.

Second, our current models only used time (the number of days) to predict subsidence. This is a simplified approach, as real-world subsidence is more complex. We did not include other important factors that can affect subsidence, such as changes in the building's load, the specific type of soil and its properties, or environmental conditions like groundwater levels and weather. Because we only used time and subsidence measurements, our models might not be as accurate or reliable in different real-world situations where these other factors are important. Therefore, an important next step for future research is to create models that use multiple features. This means we should try to include more types of data, like soil reports, information on building loads, and groundwater data, to develop more complete and likely more accurate prediction models.

Third, the irregular measurement frequency (ranging from 20 days to over 9 months) challenges the models in accurately predicting distant data points, particularly for LR and GB, as evidenced by larger errors in the September 29, 2022, cycle.

Fourth, the validation of our models was based on a very small test set, comprising only the last three measurement cycles. This limited number of test data points increases the risk that overfitting to the training data. To address this in future studies and ensure more robust model evaluation, more rigorous validation techniques should be employed. Specifically, k-fold cross-validation would be highly beneficial, especially with limited datasets, as it allows for all data points to be used for both training and validation. Alternatively, for time-series data like ours, specialized time-series cross-validation methods (e.g., walk-forward validation or rolling-origin cross-validation) should be considered to respect the temporal dependencies in the data and provide a more reliable assessment of future predictive performance.

Additionally, the study focuses solely on a single high-rise building, which may not represent other structure types, such as bridges or dams. These limitations highlight the need for future research to address these gaps and enhance the applicability of ML models in predicting structural subsidence.

5. Conclusion

This study analysis compared four machine learning models (LR, SVM, RF, and GB) for predicting structural subsidence of the Dormitory–Canteen building in Hanoi, utilizing 11 cycles of leveling data. RF emerged as the most optimal model, achieving a mean absolute error (MAE) ranging from 0.7161 to 1.3167 mm and a coefficient of determination (R^2) between 0.8723 and 0.8876 across all points, outperforming the other evaluated methods. This outcome underscores RF's ability to capture nonlinear subsidence trends, even with a small dataset and irregular measurement frequencies. LR, despite its simplicity, delivered acceptable performance, making it suitable for rapid-deployment applications. In contrast, GB and SVM fell short of expectations, likely due to data size limitations. These findings contribute significantly to structural subsidence monitoring, particularly within the Vietnamese context where periodic leveling-based monitoring is mandatory. The demonstrated efficacy of RF enhances the accuracy of subsidence prediction and offers a flexible tool for handling the challenges of real-world field data, such as noise and inconsistent frequencies. LR, with its stable performance, remains a viable alternative when computational resources are limited, paving the way for modernizing subsidence monitoring processes.

However, the study reveals limitations that warrant future attention. The small dataset size (11 cycles) impacted the performance of GB and SVM while constraining the models' generalizability. Additionally, the absence of data on external factors (e.g., loads, geology, environmental conditions) may diminish the practical utility of predictions in complex scenarios. This work contributes to addressing the data scarcity challenge in developing nations by evaluating the robustness of ML models under minimal and irregular datasets.

Building upon these findings, several specific next steps are recommended for future research to further advance subsidence prediction capabilities. Firstly, future models should aim to integrate external geotechnical features, such as soil moisture content, load variations on the structure, and detailed geological data, which could significantly enhance predictive accuracy by providing a more holistic understanding of subsidence drivers. Secondly, to address the challenges posed by small datasets, which are common in such monitoring scenarios, the application of advanced techniques like transfer learning or specialized deep learning architectures designed for limited data environments should be explored. Finally, a practical extension of this work would be the development of a real-time, or near real-time, monitoring and prediction tool based on the well-performing RF model, potentially incorporating an adaptive learning mechanism to continuously refine predictions as new leveling data becomes available.

Acknowledgement

The authors would like to thank Hanoi University of Civil Engineering. We are grateful to all of those with whom we have had the pleasure to work with during this and other related projects.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Nguyen Dinh Huy, Tran Dinh Trong;* ***data collection:*** *Luong Ngoc Dung, Tran Thi Hue;* ***analysis and interpretation of results:*** *Nguyen Dinh Huy, Luong Ngoc Dung, Tran Dinh Trong, Tran Thi Hue;* ***draft manuscript preparation:*** *Luong Ngoc Dung, Tran Dinh Trong, Nguyen Dinh Huy. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] J. A. Charles and H. D. Skinner, "Settlement and tilt of low-rise buildings," in *Proceedings of the Institution of Civil Engineers - Geotechnical Engineering*, Apr. 2004, pp. 65–75. doi: 10.1680/geng.2004.157.2.65.

[2] K. Karila, M. Karjalainen, J. Hyyppä, J. Koskinen, V. Saaranen, and P. Rouhiainen, "A Comparison of Precise Leveling and Persistent Scatterer SAR Interferometry for Building Subsidence Rate Measurement," *ISPRS Int. J. Geo-Information*, vol. 2, no. 3, pp. 797–816, Aug. 2013, doi: 10.3390/ijgi2030797.

[3] A. Serrano-Juan, E. Pujades, E. Vázquez-Suñè, M. Crosetto, and M. Cuevas-González, "Leveling vs. InSAR in urban underground construction monitoring: Pros and cons. Case of la sagrera railway station (Barcelona, Spain)," *Eng. Geol.*, vol. 218, pp. 1–11, Feb. 2017, doi: 10.1016/j.enggeo.2016.12.016.

[4] TCVN 9360, "Technical process of settlement monitoring of civil and industrial building by geometrical levelling." Ministry of Construction, Vietnam, 2012.

[5] TCVN 9363, "Building surveys - Geotechnical investigation for high rise building." Ministry of Construction, Vietnam, 2012.

[6] TCVN 9364, "High - rise buildings - Technical guide for survey work during construction." Ministry of Construction, Vietnam, 2012.

[7] T. Hoc Quang, N. Le Thanh, and T. Thi Hanh, "Study and establish subsidence forecast models in accordance with analysis and forecast soft ground subsidence from monitoring results," *J. Min. Earth Sci.*, vol. 58, no. 4, pp. 93–100, 2017.

[8] D. T. Tran, N. D. Luong, and D. H. Nguyen, "Prediction of building subsidence in Vietnam using machine learning techniques based on leveling results," *Geod. Cartogr.*, vol. 50, no. 3, pp. 150–155, Dec. 2024, doi: 10.3846/gac.2024.20237.

[9] C. Liu and J. Macedo, "Machine learning-based models for estimating liquefaction-induced building settlements," *Soil Dyn. Earthq. Eng.*, vol. 182, p. 108673, Jul. 2024, doi: 10.1016/j.soildyn.2024.108673.

[10] J. Zhang *et al.*, "Urban ground subsidence monitoring and prediction using time-series InSAR and machine learning approaches: a case study of Tianjin, China," *Environ. Earth Sci.*, vol. 83, no. 16, p. 473, Aug. 2024, doi: 10.1007/s12665-024-11778-w.

[11] M. Khajehzadeh, S. Keawsawasvong, V. Kamchoom, C. Shi, and A. Khajehzadeh, "Developing effective optimized machine learning approaches for settlement prediction of shallow foundation," *Heliyon*, vol. 10, no. 17, p. e36714, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36714.

[12] Y. Xu, Z. Wu, H. Zhang, J. Liu, and Z. Jing, "Land Subsidence Monitoring and Building Risk Assessment Using InSAR and Machine Learning in a Loess Plateau City—A Case Study of Lanzhou, China," *Remote Sens.*, vol. 15, no. 11, p. 2851, May 2023, doi: 10.3390/rs15112851.

[13] A. Gomez-Cabrera and P. J. Escamilla-Ambrosio, "Review of Machine-Learning Techniques Applied to Structural Health Monitoring Systems for Building and Bridge Structures," *Appl. Sci.*, vol. 12, no. 21, p. 10754, Oct. 2022, doi: 10.3390/app122110754.

[14] N. Q. Long, T. Van Anh, and B. Khac Luyen, "Determination of Ground Subsidence by Sentinel-1 SAR Data (2018-2020) over Binh Duong Quarries, Vietnam," *VNU J. Sci. Earth Environ. Sci.*, vol. 37, no. 2, Jun. 2021, doi: 10.25073/2588-1094/vnuees.4605.

[15] N. Minh Hai and T. Van Anh, "Research application of the InSAR technology for determining changes in surface topography," *J. Min. Earth Sci.*, vol. 48, pp. 20–24, 2014.

[16] Q. L. Nguyen, Q. M. Nguyen, D. T. Tran, and X. N. Bui, "Prediction of ground subsidence due to underground mining through time using multilayer feed-forward artificial neural networks and back-propagation algorithm – case study at Mong Duong underground coal mine (Vietnam)," *Min. Sci. Technol.*, vol. 6, no. 4, pp. 241–251, Dec. 2021, doi: 10.17073/2500-0632-2021-4-241-251.

[17] D. Van Phong *et al.*, "Analysis of land vertical movement using ANN function from the results of processing GNSS time series data," *Vietnam J. Hydrometeorol.*, vol. 8, no. 752, pp. 41–50, Aug. 2023, doi: 10.36335/VNJHM.2022(752).41-50.

[18] P. Quoc Khanh and N. Van Manh, "Application of artificial neural network for forecasting the subsidence of hydropower structure," *J. Min. Earth Sci.*, vol. 60, no. 4, pp. 59–66, 2019.

[19] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear Regression," in *An Introduction to Statistical Learning*, 2023, pp. 69–134. doi: 10.1007/978-3-031-38747-0_3.

[20] A. Mammone, M. Turchi, and N. Cristianini, "Support vector machines," *WIREs Comput. Stat.*, vol. 1, no. 3, pp. 283–289, Nov. 2009, doi: 10.1002/wics.49.

[21] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Stat.*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.