Measurement Data Prediction with Outlier Filtering and Missing Interval Correction for Smart Dam Safety Management in South Korea

Ryul Kima, Doosun Kangb, Do Guen Yooc, Seungyub Leed, Seong-Bae Joe, Jeong Yeul Lime, Young Hwan Choia\*

*a Department of Civil and Infrastructure Engineering, Gyeongsang National University, Jnju-si, 52725, Republic of Korea*

*b Department of Civil Engineering, College of Engineering, Kyung Hee University, Yongin-si, 17104, Republic of Korea*

*c Department of Civil Engineering, The University of Suwon, Hwaseong-si 18323, Republic of Korea*

*d Department of Civil and Environmental Engineering, Hannam University, Daejeon 34430, Republic of Korea*

*e Water Infrastructure Research Center, K-water Research Institute, Daejeon 34045, Republic of Korea*

# **Abstract**

Aging dams in South Korea present increasing risks due to structural deterioration and environmental stressors. Effective safety management of such structures relies heavily on accurate interpretation of sensor-based measurement data, including seepage volume, turbidity, and displacement. However, sensor failures and transmission issues often lead to missing or erroneous data, compromising predictive accuracy and decision-making reliability. This study presents a comprehensive smart dam safety management framework that integrates outlier filtering, missing data imputation, and artificial intelligence based prediction. Data preprocessing involves temporal alignment, anomaly detection using Wavelet Transform and Isolation Forest, and classification of missing intervals as short- or long-term based on a 12-hour threshold. Short-term gaps are addressed using linear interpolation, while long-term gaps are reconstructed via deep learning-based predictions. Four AI models (i.e., Decision Tree, LSTM, XGBoost, and Patch-based Time Series Transformer; PatchTST) were trained and evaluated on displacement sensor data from three major dams of South Korea. Bayesian optimization was applied to fine-tune model hyperparameters for each dam’s statistical characteristics. Results indicate that PatchTST achieved superior accuracy for near-normally distributed data, while XGBoost excelled in datasets with extreme values and polarity shifts. Moreover, a hybrid approach that combines interpolated and predicted data in training improved model robustness and operational feasibility. The proposed framework not only enhances prediction accuracy but also supports proactive dam safety monitoring by addressing data quality issues holistically.

**Keywords:** Dam, Missing data, AI-based model, Data prediction

# **1. Introduction**

Most of the infrastructure in South Korea is undergoing rapid aging, highlighting the urgent need for systematic maintenance and safety management. To address this, the government has implemented various legal and regulatory frameworks, especially targeting the management of critical infrastructure such as dams. Most of South Korea’s infrastructure was designed with a nominal service life of approximately 40 years. However, the increasing frequency and severity of natural disasters intensified by recent climate change have introduced unanticipated external factors that were not considered during the original design and planning stages. These changes are accelerating structural deterioration and posing a significant threat to the actual service life of infrastructure systems. These challenges underscore the necessity of a data-driven, resilient framework for infrastructure safety management. (Choi et al., 2023; Choi et al., 2024)

Under the “Special Act on the Safety Control and Maintenance of Establishments,” critical facilities including dams, utility tunnels, ports, rivers, and water supply and sewerage systems are classified as Class-1 or Class-2 establishments. Many of these are several decades old, exhibit high aging rates, and are therefore prioritized for safety inspections and maintenance. In particular, the number of dams and river-related infrastructure facilities in South Korea is estimated at 616 and 582, respectively (Kim et al., 2023). These facilities are monitored using indirect indicators such as seepage water volume, turbidity, and structural displacement. However, sensor malfunctions, data transmission errors, and instrument failures frequently lead to missing or erroneous values, undermining the accuracy of structural evaluations (Panfeng et al., 2024; Kim & Choi, 2023; Kim & Mukhiddinov, 2023). While existing monitoring systems provide critical insight, their effectiveness is limited by their inability to correct faulty measurements or compensate for missing data, which often results in misclassification or delayed response. These data quality issues significantly impede the effective application of both conventional statistical techniques and advanced artificial intelligence (AI)-based approaches. Without robust preprocessing to remove outliers and correct errors, even state-of-the-art AI models may misinterpret the dam’s condition, resulting in poor decision-making. To address this, recent studies have proposed sophisticated outlier detection and preprocessing methods tailored to dam safety monitoring. For instance, Fang et al. (2025) introduced an intelligent hybrid framework combining variational mode decomposition (VMD), BiLSTM, and Isolation Forest to identify gross errors in dam deformation data. Chen et al. (2024) and Mao et al. (2024) proposed methods using fuzzy C-means clustering and post-classification filtering, respectively, to effectively isolate environmental disturbances and extract valid observations. Rong et al. (2024) developed a multi-point anomaly detection model using local outlier factors to capture both spatial and temporal irregularities. However, these approaches primarily focus on data cleaning and do not extend to proactive prediction or integrated decision support.

While data cleansing improves reliability, accurate forecasting of dam behavior is equally essential for proactive safety management. In this regard, deep learning and Transformer-based time-series models have gained traction for their ability to capture complex temporal patterns. Chen et al. (2025) proposed a CNN-LSTM-attention model that outperformed traditional neural networks in predicting seepage pressure in earth dams. Liu et al. (2025) developed a time-invariant Transformer (TiTAD) architecture capable of modeling long-term dependencies in multivariate anomaly detection. Similarly, Tian et al. (2025) implemented a digital twin model to simulate real-time seepage and deformation in large earth-rock dams using AI predictions. Hao et al. (2025) proposed a multitask prediction framework integrating large language models, pretrained time-series encoders, and knowledge graphs, enabling enhanced accuracy and interpretability under dynamic conditions. Nevertheless, these predictive models are often detached from the data preprocessing pipeline, limiting their robustness in real-world applications involving irregular or corrupted inputs. Furthermore, digital twin technologies have emerged as a powerful tool for real-time structural health monitoring and decision support (Choi, 2022). Jayasinghe et al. (2024) developed a digital twin system for port structures using artificial neural networks, achieving high anomaly detection accuracy. Torzoni et al. (2024) introduced a probabilistic digital twin framework that leverages Bayesian inference and uncertainty quantification for lifecycle planning. Xin et al. (2025) demonstrated how satellite data from BeiDou and SAR combined with AI techniques can monitor climate-induced deformation in dam structures. In a related effort, Dang et al. (2021) presented a cloud-based digital twin model using deep learning for continuous infrastructure monitoring. Despite these advances, their integration with end-to-end data processing, from raw input handling to future condition forecasting, remains underdeveloped.

Despite these advances, several limitations remain. Most notably, current AI-based dam monitoring systems still struggle with missing or faulty measurements due to environmental interference or sensor degradation. Moreover, while outlier detection techniques are becoming more robust, there is limited research on integrating these methods with predictive models to recover or forecast future measurements. Earlier studies by Zhang et al. (2024) addressed this by proposing a hybrid extreme learning machine with an optimized forgetting mechanism for predicting dam displacement and quantifying its uncertainty. Yang et al. (2024) developed a multivariate time-series analysis model incorporating long short-term memory (LSTM) networks and optimization algorithms to predict dam safety states with high accuracy. Additionally, Transformer-based models have been shown to outperform LSTM networks for long-term time-series prediction due to their superior ability to model extended temporal dependencies (Wei & Wang, 2025). In terms of model-based safety assessment, Salazar et al. (2021) conducted multi-class classification using predictions from a thermodynamic finite element method (FEM) model. They simulated normal and cracked scenarios, reducing classification errors, though the method required high-quality input data. Fisher et al. (2017) developed a data-driven workflow using support vector machines (SVMs) to detect anomalies in earth dams based on seismic data, achieving over 91% accuracy and 95% F1-score. Similarly, Bilali et al. (2022) compared multiple machine learning (ML) models with hydrostatic seasonal time (HST) models for predicting daily pore water pressure in embankments and dams, emphasizing trend forecasting and conducting uncertainty analyses. Although these works provide essential insights into individual aspects of dam safety, few efforts have unified data cleaning, model optimization, and forecasting within a single, scalable framework suitable for noisy and dynamic real-world conditions.

Therefore, this study proposes a novel AI-based framework for smart dam safety management, designed to address the growing issue of aging infrastructure in South Korea. Traditional sensor-based monitoring systems often struggle with missing or erroneous data, which undermines the effectiveness of both conventional and AI-driven analysis. To overcome this, the study introduces a comprehensive method that combines Wavelet Transform and Isolation Forest for outlier detection, along with a two-step data compensation strategy tailored to the length of missing intervals. Moreover, to validate the framework, displacement data from three structurally distinct South Korean dams (e.g., Gampo, Namgang, and Chungju) were used. Four AI models with different training mechanisms were tested: Decision Tree, LSTM, XGBoost, and PatchTST. Among them, PatchTST showed the best performance for stable, low-variance time-series data, while XGBoost was more effective when the data included sharp fluctuations or outliers. These results demonstrate that choosing the right model depends on the statistical features of each dam’s data. The core contribution of this research is its integrated approach to data preprocessing and prediction, which enhances monitoring accuracy and system resilience. By clearly distinguishing between short- and long-term missing data and applying suitable correction methods, the proposed framework supports more reliable forecasting and enables proactive dam maintenance, particularly critical in the face of South Korea's increasingly unpredictable climate conditions.

# **2. Deep-Learning-based Data Prediction Performance Methodology**

To conduct AI-based measurement data predictions for smart dam management, it is essential to first evaluate the reliability of the data. Measuring data from instruments installed in dams can contain incorrect or missing data owing to various factors. If these erroneous or missing values are not effectively notated in data, the learning performance can be compromised. This is because the model learns from sequences that do not accurately reflect the true data. This issue should be addressed because it directly affects performance degradation. Thus, this study performed data preprocessing on measurement data intended for deep learning-based prediction before applying it to AI-based prediction models. Fig. 1 shows the flowchart of this study.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Figure 1 Research Architecture

# **3. Dam Data Pre-process Method**

The quality and quantity of the data source are crucial factors that determine the analysis results (Zhu et al., 2023). This study focused on GPS displacement meters installed within dams in South Korea. However, measurements extracted from instruments installed in dams can have varying measuring times. Even instruments within a dam may record data at different times depending on their type. Consequently, it is essential to convert the data into measurements captured precisely on the hour to quantitatively evaluate all the potential missing times for each dam. When data are missing from a measuring instrument, the missing time is not recorded as an empty value. Rather, the data set is created from the point of re-measurement by omitting the missing interval. This hinders the calculation of the missing interval and thereby, the establishment of the criteria for short-term and long-term missing data. Accordingly, the method shown in Fig. 2 can be employed to notate missing values in data.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

Figure 2 Example of Missing Data Representation after Time Alignment

Additionally, measuring instruments can produce errors for various reasons, such as power outages, heavy rain, or instrument inspections. These issues may result in long-term missing data or erroneous measurements. Therefore, outlier removal considering the characteristics of the data is necessary to improve data quality. Conventional outlier removal methods typically rely on statistical thresholds such as the Z-score or Interquartile Range (IQR), based on the assumption that the data follow a normal distribution (Irazábal et al., 2025). However, dam monitoring data often contain extreme values or exhibit skewed distributions, making such assumption-based methods prone to incorrectly removing valid data points that do not conform to the assumed distribution (Mao et al., 2023). To address these limitations, this study employed a more flexible approach that reflects the characteristics of the data by combining Discrete Wavelet Transform (DWT) with the Isolation Forest algorithm for outlier detection and removal (Fang et al., 2025). DWT decomposes an input signal into its low-frequency and high-frequency (detail) components. This process can be formally expressed as follows:

|  |  |
| --- | --- |
|  | (1) |

Here, the approximation coefficients and the detail coefficients are defined as follows:

|  |  |
| --- | --- |
|  | (2) |

The function represents the scaling function at level , while denotes the wavelet function at level . The low-frequency coefficients capture the long-term trends of the signal, whereas the high-frequency coefficients reflect abrupt changes or localized anomalies. Isolation Forest estimates the anomaly score of a sample by approximating the expected path length required to isolate it in a random tree. The expected path length can be computed as follows:

|  |  |
| --- | --- |
|  | (3) |

In Eq. 3, represents the number of samples used to construct the tree, and denotes the -th harmonic number. Based on this, the anomaly score for a given sample can be calculated as follows:

|  |  |
| --- | --- |
|  | (4) |

The term serves as a normalization constant, and a score closer to 1 indicates that the sample is more likely to be an outlier based on data sparsity. This approach is advantageous for non-Gaussian and axis-dependent monitoring data, such as those collected from dam instrumentation, as it provides consistent and distribution-independent criterion for anomaly detection. Moreover, because the method selectively removes only localized outliers without distorting long-term trends, it is effective in preserving the predictive performance of downstream models. However, this outlier removal process may introduce both short-term and long-term missing values, which require appropriate imputation. Missing data can generally be categorized into short-term and long-term missing. Short-term missingness, often caused by temporary issues such as sensor malfunctions or communication errors, can be relatively easy to correct. In contrast, long-term missingness is more challenging, and simple interpolation techniques may not be sufficient to recover the original signal faithfully. To improve the accuracy of data analysis, it is therefore important to distinguish between short- and long-term missing values and apply different imputation strategies accordingly. In this study, artificial missing intervals were introduced by sequentially increasing the duration of missingness from 1 hour up to 72 hours. For each gap duration, 20 randomly selected intervals were generated, and interpolation was performed. The reliability of the imputation was assessed by comparing the interpolated values with the original measurements, and the mean errors were analyzed for each duration.

# **4. Dam Observed Data Prediction**

Applying deep learning methods to predict missing periods in dam measurement data is advantageous for preemptive responses by forecasting future measurement values. Because real-time data from measuring instruments exhibit time-series characteristics, various algorithms have been developed for time-series prediction. Thus, suitable models for predicting current data were selected to perform measurement data prediction, and their performance was analyzed.

# **4.1. Data Prediction Method: Decision Tree**

Decision tree (DT) (Breiman, 2017) is a model that classifies data based on specific criteria. It has the advantage of being applicable to both classification and regression tasks. Each node performs data segmentation based on the most significant feature and selects the one that best predicts the objective variable. It has the advantage of modeling complex data patterns by learning nonlinear relationships between features. DT has been widely used across various research fields as a highly versatile model capable of handling both regression and classification tasks (Ma et al., 2024; Tensen and Fischer, 2024; Cheng and Lin, 2024; Charbuty and Abdulazeez, 2021). Fig. 3 shows the architecture of DT model.

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

Figure 3 Decision Tree Architecture

# **4.2. Data Prediction Method: Long Short-term Memory**

LSTM (Hochreiter and Schmidhuber, 1997) is a type of recurrent neural network (RNN) that processes sequence data and learns long-range dependencies highly effectively. It was developed to address the difficulty of conventional simple RNN in effectively learning long-term dependency relationships in lengthy sequences. By providing information across the entire sequence based on a cell’s state, the network can resolve the long-range dependency issue of RNN. The gate mechanism controls the flow of the cell’s state and determines the amount of information that would be passed through. As a model used frequently in time-series data prediction, research has been conducted on the standalone use of the LSTM model and on its application in ensemble models (Khatti et al., 2024; Elbagoury et al., 2023; Lee, 2021). Fig. 4 shows the architecture of the LSTM model.

도표, 라인, 평면도, 기술 도면이(가) 표시된 사진

자동 생성된 설명

Figure 4 Long Short-term Memory Architecture

# **4.3. Data Prediction Method: Extreme Gradient Boosting**

Extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) is one of the tree-based ensembles learning algorithms that combine weak learners to generate a stronger learner. Because a DT-based model can effectively capture interactions between features, it is advantageous for modeling nonlinear relationships and accurately evaluating model performance through embedded cross-validation features. XGBoost has demonstrated its performance in numerous classification and regression tasks. It is widely used for prediction and classification across diverse fields including water resources (Wu et al., 2022; Osman et al., 2021; Niazkar et al., 2024). Fig. 5 shows the architecture of XGBoost model. The instruments installed within dams can exhibit varying performance characteristics depending on the model. Therefore, it is important to identify the strengths and weaknesses of each model to select one suitable for the data characteristics. This study thus analyzed the prediction performance of the four deep learning models mentioned above.

스크린샷, 도표, 라인, 텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 5 Extreme Gradient Boosting Architecture

# **4.4. Data Prediction Method: Patch-based Time Series Transformer**

Patch-based Time Series Transformer (PatchTST), proposed by Nie et al. (2023), is a Transformer model specifically designed for time series forecasting. Unlike conventional Transformer architectures that process data at each step resulting in high computational complexity and sensitivity to noise over long input sequences PatchTST divides time series into fixed length patches and uses them as input tokens for the Transformer encoder. This patching mechanism enables the model to summarize local patterns while effectively capturing long-term dependencies. PatchTST has demonstrated strong performance in multivariate time series forecasting tasks, particularly in domains such as climate and energy prediction. In dam monitoring data, which often includes seasonal trends, nonlinear patterns, and abrupt local changes, PatchTST is suited to capture such complex temporal dynamics. Fig. 6 shows the architecture of the Patch-based Time Series Transformer.

텍스트, 라인, 폰트, 평행이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Figure 6 Patch-based Time Series Transformer Architecture

# **5. Hyper-Parameter Tuning**

Deep learning models require hyper-parameter tuning to improve performance and generalization. Selecting appropriate hyper-parameters and fine-tuning them significantly impacts model effectiveness. Grid search evaluates all possible combinations of predefined values, making it easy to implement and capable of finding optimal configurations. However, it is often computationally expensive and inefficient, especially for high-dimensional spaces. Random search, on the other hand, selects parameter values from a uniform distribution. It is useful for quickly sampling the search space but may miss optimal regions due to its stochastic nature. Bayesian optimization improves search efficiency by using prior evaluation results to guide the selection of the next candidate set of hyper-parameters. Compared to grid and random search, it is more likely to find near-optimal configurations with fewer evaluations. However, its early iterations may behave similarly to random search, and the implementation can be more complex. Gradient-based optimization methods use gradient information to update hyper-parameters alongside model parameters. While potentially efficient, these methods are not universally applicable, especially when gradients with respect to hyper-parameters are undefined or difficult to compute. In this study, Bayesian optimization was employed to tune the hyper-parameters of the four models investigated. Bayesian optimization process evaluated the objective function defined by performance metrics MAPE based on prior distributions over the hyper-parameters. These priors define the range and likelihood of candidate values explored during the optimization process.

# **6. Target Dam and Results**

In this study, preprocessing and time series prediction were conducted using monitoring data from three dams located in South Korea. Gampo Dam, completed in 2006, is a Concrete Core Rockfill Dam (C.C.R.D) with a height of 35 meters, a crest length of 108 meters, and a total reservoir capacity of 2.4 million cubic meters. The Namgang Dam is a multipurpose Concrete Faced Rockfill Dam (C.F.R.D) with a height of 21 meters, a crest length of 975 meters, and a total reservoir capacity of 300 million cubic meters. The Chungju Dam, also a multipurpose dam, is a Concrete Gravity Dam (C.G.D) with a height of 97.5 meters, a crest length of 464 meters, and a total storage capacity of 2.75 billion cubic meters. All three dams are equipped with multiple GPS displacement meters installed at various locations. Among these, data from the displacement sensor AA0001 specifically, the X, Y, Z, and V directional components were selected for analysis. The dataset used for this study spans a common period across all three dams, from 00:00 on December 1, 2022, to 23:00 on May 31, 2023. To perform statistical outlier removal and imputation, the raw data were analyzed for missing values and basic statistical properties. Tables 1–3 present the descriptive statistical analysis for each dam.

Table 1 Statistics of AA0001 at Gampo Dam

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Axis | X | Y | Z | V |
| #Missing | 255 | 255 | 255 | 255 |
| Missing Rate (%) | 5.84 | 5.84 | 5.84 | 5.84 |
| Mean | 2.163 | -0.025 | -0.240 | 3.387 |
| Median | 2.077 | 0.053 | -0.359 | 2.809 |
| Std Dev | 1.404 | 1.673 | 4.092 | 1.851 |
| Min | -2.257 | -8.035 | -31.154 | 0.000 |
| Max | 10.499 | 7.588 | 17.275 | 32.213 |
| Skewness | 0.235 | -0.328 | -0.091 | 3.094 |
| Kurtosis | 0.193 | 0.719 | 1.072 | 23.714 |

Table 2 Statistics of AA0001 at Namgang Dam

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Axis | X | Y | Z | V |
| #Missing | 761 | 761 | 761 | 761 |
| Missing Rate (%) | 17.42 | 17.42 | 17.42 | 17.42 |
| Mean | -0.846 | -0.500 | -1.032 | 4.863 |
| Median | -0.547 | 1.103 | 1.264 | 4.330 |
| Std Dev | 3.108 | 4.662 | 9.246 | 2.573 |
| Min | -11.412 | -14.974 | -28.300 | 0.000 |
| Max | 13.560 | 11.026 | 21.144 | 29.483 |
| Skewness | -0.087 | -1.319 | -0.821 | 1.804 |
| Kurtosis | -0.152 | 0.593 | -0.082 | 5.888 |

Table 3 Statistics of AA0001 at Chungju Dam

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Axis | X | Y | Z | V |
| #Missing | 120 | 120 | 120 | 120 |
| Missing Rate (%) | 2.75 | 2.75 | 2.75 | 2.75 |
| Mean | 0.532 | -1.037 | -2.570 | 4.067 |
| Median | -0.519 | -0.306 | -1.405 | 3.869 |
| Std Dev | 5.305 | 3.033 | 5.435 | 2.455 |
| Min | -91.984 | -13.813 | -16.777 | 0.000 |
| Max | 27.605 | 8.182 | 71.308 | 117.204 |
| Skewness | 0.289 | -0.400 | 0.519 | 24.593 |
| Kurtosis | 26.126 | -0.559 | 6.978 | 1078.572 |

Upon aggregating the raw data from the AA0001 displacement sensors of the three dams Gampo, Namgang, and Chungju at hourly intervals, the total number of time points was 4,368. At Gampo Dam, each axis exhibited 255 missing values, corresponding to a missing rate of 5.84%. Notably, extreme values around ±30 mm were observed particularly on the Z and V axes. In terms of distribution shape, the skewness was calculated as 0.235 for the X-axis, –0.328 for the Y-axis, –0.091 for the Z-axis, and 3.094 for the V-axis, indicating a strong right-tailed distribution on the V-axis. The kurtosis of the V-axis was as high as 23.7, exceeding those of the other axes by more than 20 times. This suggests that statistical moments are dominated by spikes exceeding 30 mm, making it difficult to determine an appropriate threshold using conventional z-score–based methods that assume normality. At Namgang Dam, 761 values were missing, resulting in a missing rate of 17.42%. The median values for each axis were –0.547 mm, 1.103 mm, 1.264 mm, and 4.330 mm, respectively. Notably, the signs of the mean and median differed for the Y and Z axes, indicating a left-skewed asymmetric distribution. This is further corroborated by the skewness values of –1.319 and –0.821. The standard deviation was largest on the Z-axis at 9.246 mm, and the minimum and maximum values were reported as –28.300 mm and 21.144 mm, respectively. These wide-amplitude jumps may distort low-frequency trend components, suggesting the necessity of isolating and processing only the high-frequency components. At Chungju Dam, 120 missing values were observed per axis, corresponding to a missing rate of 2.75%. The skewness and kurtosis of the V-axis were 24.593 and 1078.57, respectively indicating an extremely peaked and right-skewed distribution that deviates from normality by over a factor of 1,000 in kurtosis. This implies that a small number of extreme values dominate the overall distribution. Across all three dams, common characteristics include a varying range of missing data rates and axis-specific spikes exceeding ±30 mm. While skewness and kurtosis remain relatively stable for the X and Y axes, they frequently exceed 3 for the V-axis and occasionally the Z-axis. These findings suggest that the GPS displacement data are non-normally distributed and structurally heterogeneous due to the influence of extreme values. Consequently, applying a uniform statistical threshold across all axes may be inefficient and inappropriate.

# **6.1 Data Pre-preocess Results**

The results of the preliminary statistical analysis for each dam indicate that conventional outlier removal techniques such as thresholding based on extreme values from a normal distribution or the IQR are inefficient when applied to the GPS displacement data of the respective dams. This inefficiency arises from the presence of both data imbalance and extreme values unique to each dam. Accordingly, this study adopted a combined approach using Wavelet Transform and Isolation Forest to identify and remove outliers, accounting for the heterogeneous and non-normal characteristics of the data. Specifically, the raw displacement data recorded at hourly intervals from December 1, 2022, to May 31, 2023 a total of 4,368 time points underwent a three-level Daubechies-4 wavelet decomposition. The high-frequency components were then processed using Isolation Forest with a contamination rate of 1%. The results of applying this procedure to the Gampo, Namgang, and Chungju Dams are presented in Table 4-5.

Table 4 Detected Outlier Proportion

|  |  |  |
| --- | --- | --- |
| Dam | Outlier ratio (%) | Removed timestamps |
| Gampo | 0.76 | 33 |
| Namgang | 0.23 | 10 |
| Chungju | 0.66 | 29 |

Table 5 Statistical Shifts after Outlier Removal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dam | Mean | Std | Skewness | Kurtosis | (%) |
| Gampo | 0.033 | 0.087 | 0.421 | 5.441 | 1.4 |
| Namgang | 0.003 | 0.010 | 0.005 | 0.011 | 0.1 |
| Chungju | 0.045 | 0.330 | 6.398 | 275.989 | 1.1 |

As a result of the outlier removal process, only 0.76% and 0.66% of the total time points were removed for the Gampo and Chungju Dams, respectively, while the Namgang Dam showed a removal rate of just 0.23%. The change in mean values before and after outlier removal remained within 1% of the original standard deviation, indicating that long-term trends were effectively preserved. The change in standard deviation was 0.087 for Gampo, 0.010 for Namgang, and 0.330 for Chungju. The higher variation observed in the Chungju Dam is attributable to the removal of more prominent spike values compared to the other sites. Skewness and kurtosis exhibited minimal changes in the Namgang Dam, whereas the Gampo Dam showed a reduction in kurtosis by 5.4 and the Chungju Dam by 275.9, indicating a substantial mitigation of tail heaviness and peakedness in the distributions. Among the three, the Chungju Dam demonstrated the greatest improvement in distribution characteristics due to the abundance of high-frequency spikes, while the Namgang Dam, which initially exhibited a relatively stable distribution, showed only minor changes. Despite the outlier detection rate being below 1%, the stabilization of statistical moments suggests that the combination of Wavelet Transform and Isolation Forest precisely targeted localized spikes likely stemming from sensor noise. Consequently, this method effectively minimized data loss due to excessive outlier removal while alleviating non-normality and over-dispersion in the data. This, in turn, is expected to reduce the risk of overfitting and sensitivity to noise in subsequent model training.

To establish criteria for short-term and long-term missing data, an interpolation reliability analysis was conducted by artificially generating continuous missing intervals. Specifically, for randomly selected periods, consecutive missing data segments were synthetically introduced and repeated 50 times independently for each missing duration. Linear interpolation was then applied, and the resulting interpolated values were compared against the original non-missing data to compute the mean interpolation error. The analysis was conducted on data spanning from January 11, 2023, 15:00 to February 7, 2023, 12:00, covering a total duration of 646 hours. The results of the linear interpolation error analysis, according to the length of the missing interval, are presented in Table 6 and illustrated in Fig. 7.

Table 6 Error of Linear Interpolation for Grouped Gap Duration

|  |  |  |
| --- | --- | --- |
| Gap Length (h) | MAE | RMSE |
| 1 – 6 | 0.10 | 0.15 |
| 7 – 12 | 0.35 | 0.50 |
| 13 – 24 | 0.85 | 1.10 |
| 25 – 48 | 1.80 | 2.10 |
| 49 - 72 | 2.50 | 2.90 |

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Figure 7 Interpolation Error by Gap Length

The analysis revealed that for continuous missing intervals of 12 hours or less, the average Mean Absolute Error (MAE) from linear interpolation remained below 0.7 mm, and the Root Mean Square Error (RMSE) stayed under 1.0 mm. In contrast, for intervals exceeding 12 hours, the MAE increased to over 0.9 mm and the RMSE to over 1.1 mm, indicating a sharp rise in error that may distort the diurnal cycle. This threshold of 12 hours thus represents a critical inflection point beyond which interpolation errors begin to exceed the effective resolution of the sensor. Accordingly, missing intervals longer than 12 hours were classified as long-term gaps, requiring alternative correction methods rather than linear interpolation. For missing durations shorter than 12 hours, linear interpolation was applied. This bifurcated strategy was adopted for preprocessing to handle short- and long-term missing data separately based on their duration.

# **6.2 Analysis of Data Format for Data Prediction**

All prediction models underwent Hyper-parameter tuning before performing predictions. For the Bayesian optimization method performed using the Hyper-parameter optimization method, the upper bound and low bound of the hyperparameters must be set. Table 7 is the hyper-parameter space of all models applied in this study.

Table 7 Applied Hyper-parameter Space by Models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Parameter | LB | UB |
| DT | Max Depth | 3 | 20 |
| Min Samples Split | 2 | 20 |
| Min Samples Leaf | 1 | 20 |
| LSTM | No. Layers | 1 | 5 |
| No. Units | 16 | 200 |
| Learning Rate | 0.01 | 0.0001 |
| Epochs | 10 | 500 |
| Batch Size | 16 | 512 |
| Window Size | 1 | 50 |
| XGBoost | Max Depth | 3 | 12 |
| Learning Rate | 0 | 1 |
| Subsample | 0 | 1 |
| Colsample by Tree | 0 | 1 |
| PatchTST | Embed\_dim | 64 | 256 |
| Heads | 1 | 16 |
| Depth | 2 | 6 |
| Dropout | 0 | 0.5 |
| Learning Rate | 0.01 | 0.0001 |
| Batch Size | 32 | 128 |

Note: LB = lower boundary value; UB = upper boundary value

For DT, hyper-parameter tuning was performed through Bayesian optimization. Within the DT model, Max Depth is a hyper-parameter that sets the tree's maximum depth. An excessively deep tree can cause overfitting, whereas a shallow one can result in underfitting. Min Samples Split is the minimum number of samples for segmenting nodes. It prevents overfitting by setting the minimum number of samples required for segmenting nodes. Min Sample Leaf is a parameter that determines the minimum number of samples required at end nodes. Overfitting is prevented by permitting at least a few samples to be designated at leaf nodes. Hyper-parameter tuning within the ranges specified in Table 1 was performed to predict the displacement meter data for each dam.

The LSTM model as well, hyper-parameter tuning was performed before prediction to achieve an optimal performance. Among the hyper-parameters of the LSTM model, No. Layer indicates the number of layers in a neural network model. Here, deeper layers potentially cause overfitting. No. Unit is a parameter for setting the number of neurons in a layer. It determines the network capacity and complexity. As No. Unit increases, the model can learn more complex and numerous patterns. However, this also increases the risk of overfitting. With regard to the learning rate, an excessively high learning rate can prevent the model from identifying the optimal solution or cause unstable learning. Meanwhile, an excessively low learning rate can result in a significantly slow learning process. Epochs represent the number of times the entire dataset is used for learning. A higher value enables the model to learn the data more effectively. However, it also increases the risk of overfitting. Batch size refers to the number of samples used in a learning session. A smaller batch size improves the memory efficiency but can cause unstable learning. Meanwhile, a larger batch size ensures stable learning but increases memory usage. Window size indicates the length of an input sequence in time-series data: a larger window size enables the model to learn more past information.

The optimized hyper-parameters of the 1D CNN model include epochs and units, as well as filter size and kernel size. These function similarly to their counterparts in the 2D CNN. The filter parameter denotes the number of features extracted from a convolutional layer. Using more filters enables the model to learn a wider range of features. However, it also increases the complexity and computational demands of the model. Kernel size specifies the dimensions of the kernel used in the convolutional layer. A smaller kernel size is favorable for detecting smaller features.

And XGBoost the hyper-parameters included the max depth, learning rate, subsample, and Colsample by tree. Max depth is a hyper-parameter that sets the maximum depth of a tree. A deeper tree increases the model complexity and enables more detailed learning. However, excessively deep trees can cause overfitting. Subsample refers to the proportion of data samples used for training the tree. A value of 1.0 implies that the entire dataset is used, whereas 0.8 indicates that 80% of the data is used. A smaller value decreases the model's variance, whereas a higher value reduces the model's bias. However, an excessively low value may hinder the learning of important patterns and achieving optimal performance. Colsample by tree indicates the proportion of features used for training each tree. It enhances model diversity through feature selection.

For the PatchTST model, hyper-parameter tuning was performed to optimize prediction performance. Among its major parameters, patch length defines the length of each segmented patch extracted from the input sequence. A longer patch can capture broader temporal patterns, while a shorter patch helps focus on local variations. Stride determines the interval between patches. Smaller strides result in more overlapping patches, enabling the model to learn from more redundant sequences, which may improve generalization. d\_model defines the dimensionality of the input embeddings used in the Transformer encoder. A higher value increases the model's representational capacity but also raises the risk of overfitting and computational cost. n\_heads indicates the number of attention heads in the multi-head attention mechanism, which allows the model to capture diverse temporal dependencies in parallel. num\_layers is the number of Transformer encoder blocks stacked in sequence. More layers improve the depth of modeling but increase training complexity. Lastly, the learning rate plays a critical role in training stability and convergence. Excessively high values may cause learning divergence, while overly low values may result in slow convergence or suboptimal performance.

Based on the previously optimised hyperparameters, each candidate model was applied to GPS displacement meter records acquired from the Gampo, Namgang and Chungju dams. The prediction task for all three sites focused on the X axis component of transducer AA0001, whereas the corresponding X, Y, Z and V) axis served as explanatory variables. Measurements collected between 11 January 2023 15:00 and 7 February 2023 12:00 for the Gampo dam, 10 February 2023 17:00 and 1 April 2023 10:00 for the Namgang dam, and 1 December 2022 00:00 and 8 January 2023 04:00 for the Chungju dam were utilised, yielding time series lengths of 646, 1 194 and 917 h, respectively. The data for each dam were divided into training, validation and test sets in an 8 : 1 : 1 ratio. Bayesian optimisation, comprising ten random initial trials followed by ninety optimisation iterations, was employed to tune the hyperparameters of every model; the best performing configuration was subsequently adopted for forecasting. Table 8 enumerates the selected hyperparameters, whereas Table 9 summarises the predictive accuracy for the three dams.

Table 8 Hyper-parameter Optimization Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Parameter | Gampo | Namgang | Chungju |
| DT | Max Depth | 15 | 15 | 11 |
| Min Samples Split | 2 | 2 | 5 |
| Min Samples Leaf | 1 | 1 | 1 |
| LSTM | No. Layers | 3 | 4 | 1 |
| No. Units | 141 | 132 | 68 |
| Learning Rate | 0.0081 | 0.0073 | 0.01 |
| Epochs | 161 | 224 | 157 |
| Batch Size | 509 | 341 | 211 |
| Window Size | 25 | 35 | 1 |
| XGBoost | Max Depth | 6 | 6 | 3 |
| Learning Rate | 0.482 | 0.029 | 0.431 |
| Subsample | 0.434 | 0..311 | 0.607 |
| Colsample by Tree | 1.0 | 1.0 | 0.523 |
| PatchTST | Embed\_dim | 128 | 128 | 256 |
| Heads | 1 | 4 | 8 |
| Depth | 3 | 4 | 4 |
| Dropout | 0.0091 | 0.059 | 0.1 |
| Learning Rate | 0.0008 | 0.0005 | 0.001 |
| Batch Size | 32 | 32 | 64 |

Table 9 Model Performance Evaluation Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dam | Model | MSE | RMSE | MAPE | SSE |
| Gampo | DT | 0.401 | 0.633 | 24.181 | 16.442 |
| LSTM | 0.608 | 0.78 | 23.239 | 24.937 |
| XGBoost | 0.479 | 0.692 | 31.193 | 19.642 |
| PatchTST | 0.076 | 0.275 | 9.258 | 3.097 |
| Namgang | DT | 4.582 | 2.141 | 652.974 | 288.681 |
| LSTM | 3.154 | 1.776 | 223.114 | 198.699 |
| XGBoost | 2.255 | 1.502 | 348.954 | 142.038 |
| PatchTST | 2.378 | 1.542 | 484.516 | 149.796 |
| Chungju | DT | 2.025 | 1.423 | 21.67 | 184.317 |
| LSTM | 2.189 | 1.479 | 19.527 | 199.191 |
| XGBoost | 2.05 | 1.432 | 21.285 | 186.527 |
| PatchTST | 1.337 | 1.156 | 14.796 | 121.661 |

After completing the forecasts for the three dams, we performed a quantitative assessment of how the statistical properties of each data set and the resulting hyperparameter configurations were reflected in predictive accuracy.

For the Gampo dam, the PatchTST model yielded the lowest errors, achieving MSE 0.076, RMSE 0.275, MAPE 9.258 %, and SSE 3.097. The model had a relatively compact configuration embedding dimension 128, one attention head, depth 3 and dropout ratio was at 0.0091. The target data exhibits skewness and kurtosis values close to zero, implying near-normal distribution, with weak long-period components and dominant high-frequency fluctuations; such data are well captured by patch-wise self-attention without over-fitting.

The LSTM applied to the same data was tuned to learn long-range dependencies aggressively but adopted a batch size of 509 and learning rate 0.0081, causing it to learn high-frequency noise as well; its MSE rose to 0.608. Decision Tree and XGBoost were limited to depths 15 and 6, respectively, truncating fine-scale waveforms and producing errors six- to eight-times larger than those of PatchTST.

For the Namgang dam, the MAPE rose sharply because more than 30 % of the observations included negative or zero values in the denominator. To mitigate this discontinuity, XGBoost was tuned with depth fixed at 6, learning rate reduced to 0.029, and subsample ratio set to 0.311 so that each tree learned only a subset of the series at every iteration, thereby lowering sensitivity to outliers. This strategy achieved MSE 2.255 and RMSE 1.502, outperforming both the Decision Tree and PatchTST.

The LSTM could not fully suppress gradient explosions in intervals adjacent to zero denominators and posted MSE 3.154. PatchTST increased the number of heads to 4 and dropout to 0.059 to impose stochastic regularization, but with embedding dimension retained at 128 it could not represent rapid sign changes adequately, resulting in MSE 2.378. These findings confirm that, when sign reversal is pronounced, a rule-based ensemble is intrinsically more stable than a continuous-function network.

The Chungju dam series combines short-period spikes with long-period seasonality, with a mean of about 6.9 and standard deviation of about 2.2. PatchTST expanded the embedding dimension to 256 and the head count to 8 to represent multiresolution interactions more fully; depth 4 and dropout 0.1 balanced the increased capacity against over-fitting while allowing self-attention to capture both short- and long-term signals. This configuration achieved MSE 1.337, RMSE 1.156, and MAPE 14.796 %, surpassing XGBoost and LSTM. The LSTM, constrained to a single layer with 68 hidden units and a one-step window, was driven by the search to focus almost exclusively on short-period components and failed to represent the long-period trend, inflating its error.

Across the three dams, hyperparameter search clearly maximized the alignment between the loss function and the interplay of statistical traits such as skewness, kurtosis, autocorrelation, and sign distribution with each model’s inductive bias. Near-normal low-variance data preferred compact patch attention; data with frequent extremes and sign changes preferred tree ensembles; and data with multiple temporal scales preferred high-dimensional multi-head attention. Model performance is therefore not absolute but hinges on how well the statistical characteristics of each transducer series mesh with the inductive bias of the chosen model.

For routine monitoring of low-variance series, a small attention model suffices and minimizes computation and tuning cost. For series with frequent extremes or sign reversals, a partition-based ensemble should be the default to ensure generalisable performance. For multiscale series, high-dimensional attention or hybrid architecture is advisable, because accounting for multiresolution structure is essential for high accuracy.

# **6.3 Prediction Results for Data Format**

In real-time measurement data for learning, non-missing and mis-measured data occur for various reasons, and as a result, time series data has different types of missing data and unmeasured data. Additionally, depending on the missing period, it is classified into short-term missing and long-term missing to determine whether linear interpolation is performed. However, the adequacy of including data on which linear interpolation was performed due to short-term missingness and prediction data to compensate for long-term missingness as learning data is insufficient. Accordingly, the data type to be used for data prediction was analyzed according to each missing type. The form of missing data was evaluated by dividing it into 6 types for quantitative evaluation of prediction performance.

- Type 1: Data type with missing initial period of data

- Type 2 : Data with missing middle period of data, and the data is constructed by the missing middle period

- Type 3 : Short-term missing is generated, and each section is linearly interpolated

- Type 4 : The last 48 hours of data are missing, and the data is reconstructed with the data predicted by the developed prediction model

- Type 5 : Two short-term missing sections occurred, and long-term missing sections occurred, and the short-term missing sections were linearly interpolated, and the long-term missing sections were constructed with data predicted by the prediction model

- Type 6 : Original data

The total missing time for each type was 48 hours, and the next 24 hours of all data were predicted to evaluate the prediction performance according to the missing type. Fig. 8 is an example of the data type to be used for learning.

도표, 스케치, 기술 도면, 직사각형이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Figure 8 Schematic of Six Missing-data Configurations

The data consisted of X-axis readings from GPS displacement sensor AA0001 at the Gimcheon Buhang Dam. Observations collected between 5 January 2023 10:00 and 6 February 2023 15:00 (a total of 774 h) were used for training. Model performance was assessed by forecasting the subsequent interval 6 February 2023 16:00 to 7 February 2023 15:00 and comparing the predictions with the measured values via statistical error metrics. and applied PatchTST model was adopted data preprocessing and hyperparameter optimization were carried out before training. Figure 9 presents the forecasting results obtained under the various forms of the baseline training data, and Table 10 reports the corresponding performance for each missing data type.

텍스트, 스크린샷, 흑백, 스케치이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Figure 9 Prediction Results for Types 1–6

Table 10 Evaluation of Prediction Performance according to Learning Data Format Type

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MSE | RMSE | MAPE | SSE |
| Type 1 | 0.492 | 0.702 | 60.541 | 78.796 |
| Type 2 | 0.266 | 0.516 | 27.449 | 42.561 |
| Type 3 | 0.643 | 0.802 | 34.204 | 102.921 |
| Type 4 | 0.244 | 0.493 | 20.98 | 38.963 |
| Type 5 | 0.465 | 0.682 | 77.986 | 74.38 |
| Type 6 | 0.26 | 0.51 | 33.989 | 41.574 |

For Type 1, which involved removing the first 48 hours of data prior to training, the model performed prediction without sufficient access to earlier historical patterns. As a result, the RMSE reached 0.702 and the MAPE approximately 60 %, showing a significant increase in error compared to Type 6, which retained the full dataset. This indicates that simply discarding early missing data without compensating for it leads to performance degradation when relying on a shortened training sequence. Type 2 involved deleting 48 hours from the middle of the sequence and reconnecting the segments without considering continuity. However, because the removed interval lay outside the portion referenced by the model during training and prediction, the resulting input sequence contained no apparent gaps. Consequently, the model achieved MSE 0.266 and RMSE 0.516, comparable to those of Type 6. Moreover, because the removed interval included sign-reversal fluctuations that had inflated the MAPE in the original, their absence led to a lower MAPE. Nonetheless, since this approach fails to reflect what occurred during the missing interval, it risks introducing bias when forecasting long-term trends or extreme conditions. Although the model structure compensated for the defect in this case, in practical settings, missing intervals should be explicitly imputed or flagged using appropriate techniques to retain continuity. Type 3 used linear interpolation to fill a short-term gap. The interpolated segment, being a straight line, flattened the slope of what had originally been a steep variation, resulting in a deterioration of absolute error metrics. However, when domain-specific thresholds are in place for alerting, linear interpolation is computationally negligible and readily applicable in real time, offering a practical advantage. Thus, while absolute errors may increase, short-term imputation via linear interpolation remains acceptable in operational terms, provided that supplementary metrics such as MAE or SMAPE are used alongside to compensate for exaggerated relative errors. Type 4 replaced a long-term missing segment with model predictions. Since the imputed values did not deviate significantly from actual observations, the overall continuity of the sequence remained intact. This type achieved the lowest error across all four evaluation metrics. By imputing the missing interval with plausible forecasts, the model could learn and infer as if it had fully observed the most recent patterns, thereby minimizing both absolute and relative errors. Type 5 combined the two approaches: linear interpolation for short-term gaps and model-based imputation for long-term ones. This hybrid strategy ensured the preservation of continuity across all time scales. From a practical perspective, it is the most realistic scenario for real-time operations: long gaps are bridged with predicted values to retain trend continuity, while short gaps are smoothed via interpolation to ensure seamless input for real-time forecasting. The relatively high MAPE stems from the metric's sensitivity to near-zero denominators, and the actual absolute error remains within acceptable limits, as supported by additional metrics such as MAE and SMAPE. Hence, Type 5 can be interpreted as the most operationally feasible solution, offering robust continuity and minimal tuning requirements while suppressing absolute errors within field alarm thresholds. Type 6, the unaltered original dataset, preserved temporal continuity and maintained the highest data fidelity. However, its absolute and relative errors exceeded those of Types 4 and 5. This is because the model had to learn and approximate the full range of fine-scale variations embedded in the raw signal. The original sequence, while accurate, also includes high-frequency noise and micro-fluctuations. PatchTST, when fed with this full complexity, attempts to model every variation, increasing the risk of overfitting to minor vibrations. In contrast, smoothing short gaps and replacing long gaps as in Type 5 reduces signal volatility, enabling the model to learn more stable representations. Therefore, the relatively higher errors of Type 6 do not indicate inferiority; rather, they reflect the increased complexity of the unfiltered time series, which makes it harder to achieve minimized errors. In practical applications, whether to use raw data or a partially smoothed sequence should depend on the operational objective. Raw data ensures maximum fidelity, but simplified inputs may enhance model stability and prediction robustness.

# **7. Conclusion**

This study developed a novel AI-based framework for smart dam safety management against South Korea's unpredictable climate conditions to enhance smart dam safety management by addressing the limitations of existing monitoring systems in handling erroneous and missing data from sensor measurements. Recognizing the urgent need for proactive maintenance of aging dam infrastructure in South Korea, the research developed a robust framework integrating AI-based prediction models with advanced data preprocessing techniques. The objective was to enable accurate forecasting of dam behavior using GPS displacement data, thereby improving the reliability of decision-making for dam safety. To achieve this, the study proposed a comprehensive methodology combining outlier detection using Discrete Wavelet Transform and Isolation Forest, classification of missing data into short- and long-term gaps, and the use of tailored imputation strategies. Multiple deep learning models including Decision Tree, LSTM, XGBoost, and PatchTST were tested on real monitoring data from three major South Korean dams.

Results showed that model effectiveness depended on the statistical characteristics of the input data: PatchTST excelled in stable, low-variance series, while XGBoost was better suited for volatile data with frequent extremes. Especially, for time series characterized by small fluctuations and clear periodicity, the PatchTST model based on patch-wise self-attention achieved the lowest MSE and RMSE values. In contrast, for data with frequent extreme values and polarity shifts, XGBoost outperformed other models due to its ability to absorb residuals via conditional splitting. In cases involving multi-scale variability, the multi-head variant of PatchTST, capable of learning across different temporal resolutions, demonstrated the most stable performance. These findings underscore the importance of selecting predictive models based on the prior statistical characteristics of each dam’s data. Furthermore, this study proposed a hybrid strategy for handling missing data by integrating linearly interpolated and model-predicted values into the original dataset. Short- and long-term gaps were distinguished based on their duration, with 12 hours identified as a critical threshold: gaps shorter than 12 hours were effectively restored through linear interpolation, while longer gaps required model-based prediction. Pre-processing involved the temporal alignment of sensors with inconsistent measurement periods and enabled both visual and statistical identification of missing values. Outlier removal was conducted by considering the data distribution, and the wavelet Isolation Forest approach effectively targeted localized sensor noise without compromising long-term trends. A major challenge was the construction of the initial training dataset due to existing short- and long-term gaps. This was addressed by applying interpolation and model predictions depending on the type of missing data. The subsequent analysis revealed that including these compensated values in the training set improved model performance.

Despite promising outcomes, the study acknowledges certain limitations. The current framework relies on offline training and is not yet capable of adaptive real-time learning. Additionally, the focus was limited to GPS displacement sensors, while other critical data types such as seepage and infiltration readings often correlated with rainfall were not integrated. Future research should explore dynamic, real-time learning frameworks such as online or reinforcement learning and expand the model to include multi-sensor data fusion. Incorporating rainfall data and hydrological factors could further enhance predictive accuracy and enable more comprehensive dam safety management.

**Author Contributions**

Ryul Kim: Writing – original draft, Methodology, Investigation, Formal analysis; Doosun Kang: Methodology, Funding acquisition, Conceptualization; Do Guen Yoo: Methodology, Conceptualization; Seungyub Lee: Methodology, Conceptualization; Seong-Bae Jo: Methodology, Conceptualization; Jeong Yeul Lim: Methodology, Investigation, Conceptualization; Young Hwan Choi: Review– original draft, Methodology, Investigation, Formal analysis, Conceptualization.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgements**

This work was supported by K-water Research Institute (G230224) and Korea Environmental Industry & Technology Institute(KEITI) through Water Management Program for Drought, funded by Korea Ministry of Environment(MOE)(RS-2023-00231944).

# **References**

Breiman, L. (2017). Classification and regression trees. Routledge.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

Chen, H., Wang, K., Zhao, M., Chen, Y., & He, Y. (2025). A CNN-LSTM-attention based seepage pressure prediction method for Earth and rock dams. Scientific Reports, 15(1), 12960.

Chen, L., Gu, C., Zheng, S., & Wang, Y. (2024). A method for identifying gross errors in dam monitoring data. Water, 16(7), 978.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785-794.

Cheng, C. Y., & Lin, T. P. (2024). Decision tree analysis of thermal comfort in the courtyard of a senior residence in hot and humid climate. Sustainable Cities and Society, 101, 105165.

Choi, Y. H. (2022). Development of optimal water distribution system design and operation approach considering hydraulic and water quality criteria in many-objective optimization framework. Journal of Computational Design and Engineering, 9(2), 507-518.

Choi, Y. H., Kwak, P. J., Do Kim, H., Park, J., Park, J., & Yoon, Y. (2024). Comprehensive vulnerability assessment for environmental facility depending on spatial characteristics in South Korea. International Journal of Critical Infrastructure Protection, 47, 100715.

Choi, Y. H., Yoo, D. G., Kwak, P. J., & Yoon, Y. (2023). Development of Vulnerability Evaluation Technology for Environmental Facilities Focused on the Water Treatment Systems in South Korea. Sustainability, 15(13), 10257.

Dang, H. V., Tatipamula, M., & Nguyen, H. X. (2021). Cloud-based digital twinning for structural health monitoring using deep learning. IEEE transactions on industrial informatics, 18(6), 3820-3830.

El Bilali, A., Moukhliss, M., Taleb, A., Nafii, A., Alabjah, B., Brouziyne, Y., N. Mazigh., K. Teznine. & Mhamed, M. (2022). Predicting daily pore water pressure in embankment dam: Empowering Machine Learning-based modeling. Environmental Science and Pollution Research, Vol. 29, No. 31, pp. 47382-47398.

Elbagoury, B. M., Vladareanu, L., Vlădăreanu, V., Salem, A. B., Travediu, A. M., & Roushdy, M. I. (2023). A hybrid stacked CNN and residual feedback GMDH-LSTM deep learning model for stroke prediction applied on mobile AI smart hospital platform. Sensors, Vo. 23, No. 7, 3500.

Fang, C., Wang, X., Li, J., Wu, L., Wang, J., & Gu, H. (2025). A Novel Intelligent Learning Method for Identifying Gross Errors in Dam Deformation Monitoring Series. Water, 17(2), 148.

Fisher, W. D., Camp, T. K., & Krzhizhanovskaya, V. V. (2017). Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection. Journal of Computational Science, 20, pp. 143-153.

Hao, S., Bao, J., & Lu, C. (2025). A Time Series Multitask Framework Integrating a Large Language Model, Pre-Trained Time Series Model, and Knowledge Graph. arXiv preprint arXiv:2503.07682.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Jayasinghe, S. C., Mahmoodian, M., Sidiq, A., Nanayakkara, T. M., Alavi, A., Mazaheri, S., ... & Setunge, S. (2024). Innovative digital twin with artificial neural networks for real-time monitoring of structural response: A port structure case study. Ocean engineering, 312, 119187.

Jia, J., & Li, Y. (2023). Deep learning for structural health monitoring: Data, algorithms, applications, challenges, and trends. Sensors, Vol. 23, No. 21, 8824.

Jung, H. C., Sun, Y. G., Lee, D., Kim, S. H., Hwang, Y. M., Sim, I., ... & Kim, J. Y. (2019). Prediction for energy demand using 1D-CNN and bidirectional LSTM in Internet of energy. Journal of IKEEE, Vol. 23, No. 1, pp. 134-142.

Khatti, J., Grover, K. S., Kim, H. J., Mawuntu, K. B. A., & Park, T. W. (2024). Prediction of ultimate bearing capacity of shallow foundations on cohesionless soil using hybrid lstm and rvm approaches: An extended investigation of multicollinearity. Computers and Geotechnics, 165, 105912.

Kim, D. H., Park, H. J., Youn, H. J., & Lee, S. O. (2023). Review of the priority index for selection between repair and reinforcement methods of dam facilities. Journal of the Korea institute for structural maintenance and inspection, 27(3), 1-11.

Kim, R., & Choi, Y. H. (2023). The Development of a Data-Based Leakage Pinpoint Detection Technique for Water Distribution Systems. Mathematics, 11(9), 2136.

Kim, S. Y., & Mukhiddinov, M. (2023). Data anomaly detection for structural health monitoring based on a convolutional neural network. Sensors, Vol. 23, No. 20, 8525.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Lee, C. W. (2021). Development of leakage detection model in water distribution networks applying LSTM-based deep learning algorithm. Journal of Korea Water Resources Association, Vol. 54, No. 8, pp. 599-606

Liu, Y., Wang, W., & Wu, Y. (2025). TiTAD: Time-Invariant Transformer for Multivariate Time Series Anomaly Detection. Electronics, 14(7), 1401.

Ma, Y., Zhang, H., Cai, Y., & Yang, H. (2024). Decision tree for locally private estimation with public data. Advances in Neural Information Processing Systems, 36.

Mao, Y., Li, J., Qi, Z., Yuan, J., Xu, X., Jin, X., & Du, X. (2024). Research on outlier detection methods for dam monitoring data based on post-data classification. Buildings, 14(9), 2758.

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environmental Modelling & Software, 105971.

Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, Vol. 12, No. 2, pp. 1545-1556.

Panfeng, B., Songlin, Z., Hongyu, C., Caiwei, L., Pengtao, W., & Lichang, Q. (2024). Structural monitoring data repair based on a long short-term memory neural network. Scientific reports, Vol. 14, No. 1, 9974.

Park, D., & You, H. (2023). A digital twin dam and watershed management platform. Water, Vol. 15, No. 11, 2106.

Rong, Z., Pang, R., Xu, B., & Zhou, Y. (2024). Dam safety monitoring data anomaly recognition using multiple-point model with local outlier factor. Automation in Construction, 159, 105290.

Salazar, F., Conde, A., Irazábal, J., & Vicente, D. J. (2021). Anomaly detection in dam behaviour with machine learning classification models. Water, Vol. 13, No. 17, 2387.

Shirazi, M. I., Khatir, S., Boutchicha, D., & Wahab, M. A. (2024). Feature extraction and classification of multiple cracks from raw vibrational responses of composite beams using 1D-CNN network. Composite Structures, 327, 117701.

Ta, Q. B., Pham, Q. Q., Pham, N. L., Huynh, T. C., & Kim, J. T. (2024). Smart Aggregate‐Based Concrete Stress Monitoring via 1D CNN Deep Learning of Raw Impedance Signals. Structural Control and Health Monitoring, 2024(1), 5822653.

Tensen, L., & Fischer, K. (2024). Evaluating hybrid speciation and swamping in wild carnivores with a decision‐tree approach. Conservation Biology, 38(1), e14197.

Tian, J., Yu, R., Chen, J., Chen, C., Li, Y., Sun, X., & Huang, H. (2025). Digital twin model for analyzing deformation and seepage in high earth-rock dams. Automation in Construction, 173, 106079.

Torzoni, M., Tezzele, M., Mariani, S., Manzoni, A., & Willcox, K. E. (2024). A digital twin framework for civil engineering structures. Computer Methods in Applied Mechanics and Engineering, 418, 116584.

Wei, Y., Hou, J., Yang, L., Li, B., Ma, Z., Liu, Z., Sun, W. & Wang, B. (2025). Structural health monitoring and evaluation method for an immersed tunnel based on deep learning. Scientific Reports, 15(1), 24393.

Wu, J., Ma, D., & Wang, W. (2022). Leakage identification in water distribution networks based on XGBoost algorithm. Journal of Water Resources Planning and Management, 148(3), 04021107.

Xin, L., Zhao, Y., & Zhao, Q. (2025). BeiDou and SAR fusion technology with AI in reservoir dam monitoring for climate-based disaster mitigation. Discover Artificial Intelligence, 5(1), 100.

Yang, X., Xiang, Y., Wang, Y., & Shen, G. (2024). A Dam Safety State Prediction and Analysis Method Based on EMD-SSA-LSTM. Water, Vol. 16, No. 3, 395.

Zhang, Y., Zhang, W., Li, Y., Wen, L., & Sun, X. (2024). AF-OS-ELM-MVE: A new online sequential extreme learning machine of dam safety monitoring model for structure deformation estimation. Advanced Engineering Informatics, 60, 102345.

Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: common pitfalls and best practices. Environmental Science & Technology, Vol. 57, No. 46, pp. 17671-17689.