Boosting Medium-Size Reservoir Water Level Predictions using Cyclical Encoding

Sakchan Luangmaneerote
Department of Computer Technology, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin, Thailand
tsakchan@hotmail.com

Wiroj Thasana
Department of Mechanical Engineering, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin, Thailand
wiroj.th@rmuti.ac.th (corresponding author)

Received: 16 February 2025 | Revised: 24 March 2025, 7 April 2025, and 10 April 2025 | Accepted: 12 April 2025

 DOI: https://doi.org/10.48084/etasr.10595

ABSTRACT

Effective feature engineering is crucial for machine learning models to capture complex data patterns. This study explores cyclical encoding, a novel technique designed to enhance the performance of machine learning algorithms on datasets with inherent periodicity. This study addresses the challenge faced by medium-sized reservoirs that lack modern data collection equipment by applying cyclical encoding to improve predictive models. Using medium-sized reservoir data, cyclical encoding was applied to enhance the predictive capabilities of models including Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), Linear Regression (LR), and Artificial Neural Networks (ANN). The results show that cyclical encoding significantly enhances the ability of nonlinear models to capture cyclical patterns, reducing error metrics by up to 75%. However, LR showed minimal improvement due to its inherent linear limitations. Feature importance analysis identified cumulative outflow and inflow volumes as key predictors. These findings highlight the vital role of advanced feature-engineering techniques such as cyclical encoding in boosting the accuracy and robustness of nonlinear machine learning models, especially for medium-sized reservoirs without modern equipment. This study underscores its potential for broader applications in domains with periodic data, such as climate modeling and financial time series.

Keywords-feature engineering; cyclical encoding; water level prediction; medium-sized reservoir; hydrology

I. INTRODUCTION

In recent years, reservoirs around the world have experienced declining water levels due to the effects of climate change, prolonged droughts [1], and sediment accumulation [2]. Medium-sized reservoirs play a crucial role in the environmental impact by minimizing ecosystem disruption and supporting easier fish migration. These reservoirs are cost-effective, require less initial investment, and are affordable to maintain [3]. However, these challenges have significantly affected the reliability of water storage. Maintaining adequate water levels in medium reservoirs is essential to ensure reliable water supply, effective flood control, and support of local ecosystems [4]. Successful management requires careful monitoring of inflows and outflows and adaptation to evolving climatic conditions [5]. These measures are critical to maintain water availability and reduce the risks associated with drought.

For optimal reservoir operation, an accurate estimate of reservoir levels is essential and requires considering factors such as inflow, discharge, water storage, infiltration, soil water content, evaporation, relative humidity, soil moisture, temperature, the El Niño climate index, and population levels [6]. However, it is recognized that collecting all these parameters in every reservoir may not be available due to financial limitations and the specific budget available for each reservoir [7]. The evolution of feature engineering in reservoir water level prediction has advanced significantly through various methodological developments and studies over the decades. Initially, the Nash-Sutcliffe efficiency coefficient was introduced to evaluate hydrological model performance, utilizing basic observable variables such as rainfall, temperature, and historical water levels through simple correlation analysis [8]. Building on this foundation, in [9], systematic model development practices were emphasized, including meticulous feature selection and data preprocessing in neural network applications for water resource predictions.

Transitioning to more data-driven approaches, Artificial Neural Networks (ANNs) were used for streamflow forecasting in [10], incorporating temporal lagging to capture hydrological dynamics, while in [11] Support Vector Machines (SVMs) were applied to predict Lake Erie water levels using historical data between 1918 and 2001, outperforming both a neural

network (MLP) and a Seasonal Auto-Regressive (SAR) model in forecasting 3 to 12 months, with its success being attributed to its structural risk minimization and global optimization capabilities. In [12], feature engineering was further improved by applying ensemble learning and Principal Component Analysis (PCA) to reduce dimensionality, improving model interpretability and reducing complexity. In [13], the superiority of ANNs over Support Vector Regression (SVR) was demonstrated in handling complex water demand data by normalizing and including lagged inputs.

Advances continued with the integration of more sophisticated feature extraction techniques. In [14], feature engineering was enhanced in water quality assessment by selecting influential parameters and calculating a Water Quality Index (WQI) in the Improved Decision Tree Learning (IDTL) model. This approach simplified classification and improved interpretability, highlighting the role of tailored feature selection. In [15], wavelet decomposition was introduced for daily lake-level forecasting, improving accuracy by capturing various frequency components. Subsequent studies [16, 17] applied logarithmic transformations and lagged water levels, respectively, to improve flood flow and water level predictions using ANN and ARIMA models. In [18], membership functions were customized in Adaptive Neuro-Fuzzy Inference Systems (ANFIS), while in [19], wavelet transforms were combined with fuzzy logic and MLP models for effective temporal data handling.

In real-time flood forecasting, a Bayesian particle filter approach was introduced in [20], dynamically adjusting Manning's roughness coefficients in response to real-time data. This form of feature engineering, involving dynamic parameter adjustment, improved the model's predictive accuracy and robustness. The integration of remote sensing and statistical techniques marked significant progress. In [21], Landsat remote sensing data was incorporated with hydrological models using an ensemble Kalman filter, enhancing monitoring accuracy in data-scarce regions. In [5], statistical learning was used to identify critical predictors of urban reservoir levels, improving model accuracy through effective feature selection. In [22, 23], custom features were employed in Random Forest (RF) models to classify reservoir effectiveness, achieving superior accuracy and stability.

With the advent of deep learning, Long Short-Term Memory (LSTM) networks were used to automate temporal feature modeling, reducing the need for manual feature engineering [24]. In [25], the ANFIS models were improved by selecting optimal lagged reservoir levels. Recent studies have employed regression trees, SVMs, and transformer models with extensive feature engineering, significantly outperforming traditional models in predicting water levels [26]. In [27], an LSTM-GRU model was proposed for water-level prediction, showing that ASOS-integrated data improve flood prediction with high accuracy, reinforcing the role of deep learning in disaster preparedness. Comprehensive reviews, such as [6], have highlighted the critical role of feature engineering techniques such as transformations, selection, and scaling in improving predictive power, with hybrid models demonstrating increased efficiency in water level prediction.

Previous studies have focused on complex or hybrid models using multiple attributes that are often resource-intensive, time-consuming, and require modern equipment. These are resources that medium-sized reservoirs typically lack, highlighting the need for accessible and efficient forecasting methods. Recent advances emphasize feature engineering, particularly cyclical encoding, which captures seasonal patterns in reservoir levels by transforming time-based features into formats that maintain their continuity. Unlike traditional numerical encoding, which can misrepresent relationships between time-based features (e.g., January and December appear numerically distant), cyclical encoding applies sine and cosine transformations to capture periodicity effectively. Studies such as [28, 29] have demonstrated its effectiveness in improving model performance in forecasting by ensuring that machine learning models recognize natural periodic patterns. This study integrates cyclical encoding to enhance the predictive power of machine learning models while maintaining model simplicity and accessibility for medium-sized reservoirs with limited data availability.

II. METHOD

This study used data from Huai Saneng, which derives its name from the local Khmer word 'Saneng.' The Royal Irrigation Department designed the Huai Saneng Irrigation Project, Surin's first, in 1952, primarily for water retention through reservoirs and dams. Figure 1 shows the location of Huai Saneng in Surin Province, while Table I presents its physical characteristics, including reservoir area, storage capacity, and elevation. Established in 1984, the Surin Irrigation Project operates at Huai Saneng Reservoir, which is located about 7 km from Surin City. Completed in 1978, the earthen dam has a height of 20 meters and a storage capacity of 20.8 million m^3. It connects to the Ampil Reservoir to manage floods and support agriculture in the area.

Fig. 1. Huai Saneng, Surin Province, Thailand.

TABLE I. PHYSICAL CHARACTERISTICS OF HUAI SANENG

Attribute	Value
Location	Surin Province, Thailand
Reservoir area	10.4 km
Storage capacity	20 million m^3 of water at full capacity
Elevation	137 m above mean sea level
Primary purpose	Irrigation, fisheries, local water supply, and tourism

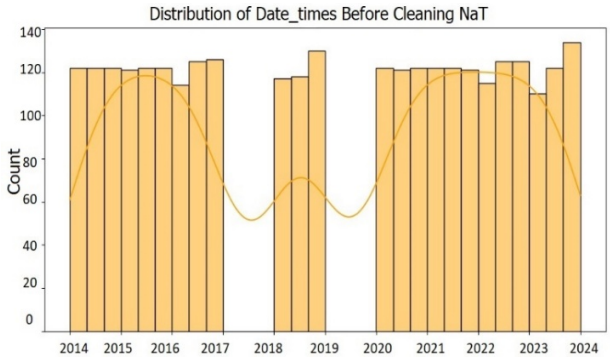

Fig. 2. Distribution of Datetimes before cleaning.

A. Data Collection and Preparation

The dataset, derived from the physical characteristics of Huai Saneng in Surin Province, Thailand, spans 8 years from 2014 to 2023, excluding 2017 and 2019. Figure 2 shows a steady count of more than 100 entries per year, except for 2017 and 2019 when data was lost due to hardware problems. However, the data for the other years are large and consistent enough to support reliable conclusions. The data includes key hydrological and meteorological attributes, essential for reservoir analysis, including Datetime, representing the recorded dates in DD/MM/YYYY format, and Capacity, indicating the total storage capacity of the reservoir in million m^3 (MCM). The Minimum_amount_of_RNOL reflects the minimum level at which no overflow occurs, measured in m, while Reservoir_storage_volume provides the current volume of stored water (in MCM). Additionally, Percent_of_RNOL specifies the percentage of capacity relative to the no overflow threshold. The dataset further includes inflow_volume (volume of incoming water expressed, MCM), Cumulative_inflow_volume (total inflow to date, in MCM), Outflow (released water volume, in MCM), and Cumulative_outflow (total outflow to date, in MCM). The usable_water_volume details the volume of water available for use (in MCM), while Rain records the amount of rainfall received (in mm), and Evaporation accounts for water loss due to evaporation (in mm). Lastly, Wind_current measures wind speed, recorded in either m/s or km/h.

The dataset comprises historical records encompassing variables such as reservoir inflow and outflow, rainfall, evaporation, wind speed, and the target variable, which is the reservoir level represented as a percentage of the reservoir's Normal Operating Level (Percent_of_RNOL). Data preprocessing procedures were employed to extract pertinent date-based features and transform date fields into a format conducive to analysis.

B. Data Cleaning

This is a pivotal preprocessing phase to ensure the integrity and reliability of the dataset. Initially, a comprehensive examination was performed to identify and quantify missing values across all variables. For numerical columns exhibiting missing entries, median imputation was employed to preserve the central tendency while mitigating the influence of outliers. In the case of temporal variables, any absent dates were addressed using the forward fill method, maintaining the chronological sequence that is essential for accurate temporal analyses. Furthermore, duplicate records were meticulously identified and removed to prevent redundancy and potential bias in subsequent analytical procedures. Particular attention was paid to verifying the consistency of measurement units, especially for the Wind_current variable, ensuring uniformity across all entries. This standardization is crucial to facilitate accurate comparisons and enhance model performance.

In addition to rectifying missing and duplicate data, the dataset underwent rigorous outlier detection and treatment to reduce the impact of extreme values. Z-scores were calculated for numerical features, and a threshold of three standard deviations was established to identify outliers. Rather than removing these anomalous data points outright, winsorization was implemented to cap values at the 1st and 99th percentiles, reducing their influence while preserving the overall data distribution. Subsequent data validation steps were undertaken to confirm that all values, such as Percent_of_RNOL, resided within logical and expected ranges (0 to 100). This comprehensive data-cleaning method established a robust foundation for subsequent feature engineering and modeling efforts, enhancing the accuracy and generalizability of the analytical results, as shown in Table II.

TABLE II. COMPARATIVE STATISTICAL SUMMARY OF FEATURES BEFORE AND AFTER WINSORIZATION

Feature	Mean_before	Mean_after	Std_before	Std_after	Skewness_before	Skewness_after	Kurtosis_before	Kurtosis_after
Capacity	22.0	22.0	7.57E-13	7.57E-13	0.00	0.00	0.00	0.00
Minimum_amount_of_RNOL	0.8	0.8	6.11E-15	6.11E-15	0.00	0.00	0.00	0.00
Reservoir_Storage_Volume	14.8	14.8	6.66E+00	6.64E+00	0.27	0.26	-0.91	-0.95
Percent_of_RNOL	67.4	67.4	3.03E+01	3.02E+01	0.27	0.26	-0.91	-0.95
inflow_volume	0.3	0.3	1.35E+00	7.54E-01	9.23	4.56	104.18	23.45
cumulative_inflow_volume	108.4	108.4	2.11E+02	2.11E+02	1.76	1.76	1.32	1.32
Outflow	0.3	0.3	1.40E+00	8.20E-01	9.14	5.24	99.58	30.02
Cumulative_outflow	120.0	120.0	2.31E+02	2.31E+02	1.73	1.73	1.21	1.21
Usable_water_volume	14.0	14.0	6.66E+00	6.64E+00	0.27	0.26	-0.91	-0.95
Rain	4.1	3.9	1.20E+01	1.07E+01	4.60	3.63	27.10	13.87
Evaporation	5.5	5.3	6.76E+00	3.12E+00	13.83	0.38	256.99	11.72
Wind_current	36.6	35.4	4.24E+01	3.40E+01	4.54	2.13	35.30	6.06

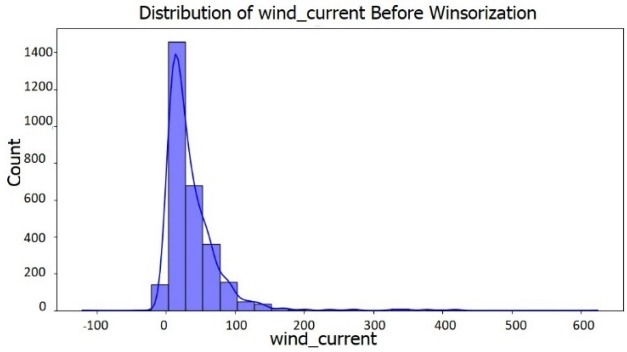

Fig. 3. Distribution of Wind_current before winsorization.

The comparative statistical summary highlights that winsorization had a targeted and effective impact on specific features. Capacity, Minimum_ amount_of_RNOL, Cumulative_inflow_volume, Cumulative_outflow, and Usable_water_volume remained entirely unchanged, suggesting that these variables were already well-behaved with no significant outliers necessitating capping. In contrast, features such as Inflow_volume, Outflow, Rain, Evaporation, and Wind_current exhibited substantial reductions in mean, standard deviation, skewness, and kurtosis. These changes indicate that winsorization successfully mitigated extreme values, resulting in more symmetric and less peaked distributions. This normalization enhances the robustness of the data, making it more suitable for predictive modeling by reducing the influence of outliers and promoting more stable statistical properties, as shown in Figures 3 and 4.

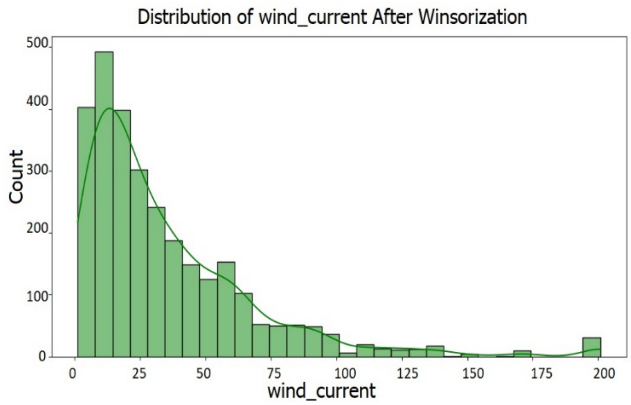

Fig. 4. Distribution of Wind_current after winsorization.

Overall, winsorization streamlined the dataset by preserving clean features and refining those with problematic outliers. This balanced approach ensures that the data maintain their integrity while improving their suitability for analytical and modeling purposes. Moving forward, these adjustments are likely to enhance model performance by fostering more reliable and generalizable predictions, particularly for statistically significant features.

C. Feature Engineering and Cyclical Encoding of Temporal Features

Data preprocessing constitutes a critical component of machine learning projects, significantly influencing the effectiveness of learning models and the precision of their outputs. Data preprocessing and feature encoding and normalization are crucial to improving machine learning accuracy and optimizing predictive maintenance schedules [30].

The concept of cyclical encoding, particularly the use of sine and cosine transformations to represent periodic features in machine learning, has evolved over time [31]. This technique leverages the mathematical properties of trigonometric functions to map cyclical data onto a unit circle, ensuring that the end of one cycle seamlessly connects to the beginning of the next. Although the exact first mention in the context of machine learning is not well-documented [32], the method has been widely adopted in various applications to address the limitations of traditional encoding methods such as one-hot encoding.

To effectively capture temporal patterns, additional features were derived from the date column, including year, month, day, day of the week, and an indicator for weekend occurrences. These features enhance the ability of the model to recognize seasonal and weekly fluctuations in water levels. To effectively model the cyclical characteristics of temporal variables such as months and days of the week within a dataset, it is advantageous to employ sine and cosine transformations. These mathematical transformations map categorical temporal features onto a continuous two-dimensional space, preserving the inherent periodicity of the data. Specifically, for the Month feature, the sine and cosine of the month number can be calculated to represent each month as a point on the unit circle. This approach ensures that December (Month 12) and January (Month 1) are positioned adjacently, reflecting their sequential relationship. Similarly, for the DayOfWeek feature, sine and cosine transformations facilitate the seamless transition between the end and the beginning of the week, such as from Sunday (Day 7) to Monday (Day 1). Integrating these transformed features into machine learning models can enhance the model's ability to capture and utilize the cyclical patterns present in temporal data, potentially improving predictive performance. Implementing these transformations involves applying trigonometric functions to the scaled temporal features. For the Month feature, the transformation can be mathematically expressed as:

$$Month\_Sin = sin(2pi * month/12) \quad (1)$$

$$Month\_cos = \cos(2pi * month/12) \quad (2)$$

Similarly, for the DayOfWeek feature, the transformations are defined as:

$$DayOfWeek\_Sin = \sin(2pi * DayOfWeek/7) \quad (3)$$

$$DayOfWeek\_Cos = cos(2pi * DayOfWeek/7) \quad (4)$$

D. Variable Selection

The target variable was the reservoir level, specifically Percent_of_RNOL. The predictive model utilizes a diverse set

of features to forecast the Percent_of_RNOL. The feature matrix X comprises hydrological variables such as Cumulative_inflow_volume and Cumulative_outflow, which represent the total water entering and exiting the system, respectively. Meteorological factors including Rain, Evaporation, and Wind_current were incorporated to account for precipitation, water loss through evaporation, and wind effects on water distribution. Temporal attributes were enhanced through Year, Month_Sin, Month_Cos, Day, DayOfWeek_Sin, and DayOfWeek_Cos, which capture both long-term trends and cyclical patterns in monthly and weekly cycles. Additionally, the binary variable IsWeekend distinguishes between weekend and weekday conditions, which may reflect variations in water usage or management practices.

The target variable y (Percent_of_RNOL) quantifies the proportion of rainfall contributing to runoff. By integrating these features, the model effectively captures the interaction between hydrological input, meteorological conditions, and temporal dynamics. The sine and cosine transformations of the month and day of the week enable the model to recognize periodic trends, enhancing its ability to make accurate and robust predictions. This comprehensive feature set supports an improved understanding and management of runoff processes, facilitating better water resource planning and decision-making.

E. Data Splitting

To ensure a comprehensive and robust evaluation of the predictive model, the dataset was partitioned into three distinct subsets: training, validation, and test sets. In [33], it is recommended to maintain separate validation and test sets to allow reliable hyperparameter optimization without contaminating the final performance metric. Specifically, 60% of the total data was allocated to the training set, which was used for model development and parameter estimation. Subsequently, 20% of the data were designated as the validation set, serving the purpose of tuning hyperparameters and facilitating model selection through performance assessment on unseen data during the training phase. The remaining 20% constituted the test set, which provided an unbiased evaluation of the model's predictive capabilities on entirely new data. This stratified splitting approach was adopted to mitigate overfitting, enhance the generalizability of the model, and ensure that performance metrics accurately reflect the model's ability to perform in real-world scenarios [34].

F. Model Selection and Hyperparameter Tuning

This study employed four predictive models, namely Linear Regression (LR), Random Forest (RF), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), to evaluate performance. The analysis involves a comparison of the results obtained with and without the application of cyclical encoding. These predictive models were employed because of their effectiveness in handling diverse relationships and their robustness against overfitting. To optimize performance, hyperparameter tuning was performed using grid search combined with cross-validation. For LR, the focus was on regularization techniques, specifically tuning the penalty parameters for L1 (Lasso) and L2 (Ridge) regularization. In SVM, the key hyperparameters included the kernel type (linear, polynomial, radial basis function), regularization parameter CCC, kernel coefficient γ, and margin tolerance. For ANN, the tuning process involved adjusting the number of hidden layers, the number of neurons per layer, the activation function, the learning rate, and the batch size. For RF, the hyperparameters included the number of estimators (trees), the maximum depth of each tree, the minimum number of samples required to split an internal node, and the maximum number of features considered for splitting at each node. By systematically adjusting these parameters, each model was fine-tuned to achieve optimal accuracy and generalizability, thus enhancing predictive performance on unseen data.

G. Test Performance and Feature Importance Analysis

The optimized model demonstrated strong predictive accuracy on the validation set, evaluated using MSE, RMSE, and R^2, ensuring its reliability on unseen data before final testing. The test results confirmed the generalizability of the model, with the MSE, RMSE, and R^2 metrics validating its ability to predict reservoir levels accurately. Additionally, feature importance extraction identified the features that contributed most to the model's predictions, providing valuable insights into the underlying data and highlighting the factors that significantly influenced the target variable (Percent_of_RNOL). This process facilitated a deeper understanding of the relationships within the data and prioritized features by sorting them from most to least important, ensuring easy interpretation and actionable insights to enhance the model's predictive capabilities.

III. RESULTS AND DISCUSSION

The results in Tables II and III indicate that the application of cyclical encoding substantially enhanced the performance of nonlinear models, such as RF, SVM, and GBR with an ANN. In particular, the RF model exhibited a significant increase in the R^2 values, rising from 0.6485 (validation) and 0.6585 (test) without encoding to 0.9223 and 0.9427, respectively, when cyclical encoding was applied. This improvement was accompanied by considerable reductions in error metrics, demonstrating the model's ability to better capture cyclical patterns in the data. This result indicates that the RF model demonstrated high accuracy and stability, with reduced sensitivity to sample size variations [22]. Similarly, SVM and GBR also experienced marked enhancements in performance across validation and test datasets, underscoring the effectiveness of cyclical encoding in improving their ability to model periodic relationships. The SVM model demonstrated superior predictive accuracy and generalizability compared to traditional models [11].

Conversely, LR, constrained by its inherent linearity, showed minimal improvement, highlighting its limitations in leveraging the encoded features. The application of cyclical encoding significantly enhanced the performance of nonlinear models by effectively capturing cyclical patterns in the data. Specifically, the RF model exhibited a substantial reduction in error metrics, with the Mean Squared Error (MSE) decreasing by approximately 83% on the test set, from 180.0700 before encoding to 30.2653 after encoding. Similarly, the MSE of the

SVM model was reduced by about 75%, dropping from 292.9606 to 73.8347 on the test set. These considerable reductions demonstrate the efficacy of cyclical encoding in improving the models' ability to model periodic relationships. Conversely, LR showed minimal improvement, highlighting its limitations in leveraging the encoded features.

Consider the target variable y as a nonlinear function of the cyclical variable x as shown in:

$$y = f\left(sin\left(\frac{2\pi x}{T}\right), cos\left(\frac{2\pi x}{T}\right)\right) + \varepsilon \quad (5)$$

Nonlinear models can approximate the function f effectively, capturing the periodic nature of y with respect to x. In contrast, a linear model would attempt to fit according to:

$$y = \beta 0 + \beta 1\left(\sin\left(\frac{2\pi x}{T}\right) + \beta 2\cos\left(\frac{2\pi x}{T}\right)\right) + \varepsilon \quad (6)$$

This linear combination cannot capture the cyclic peaks and troughs unless the relationship between y and the encoded features is strictly linear, which is rarely the case in cyclical data.

Cyclical encoding serves as a crucial feature-engineering technique, particularly for nonlinear models, allowing them to effectively exploit periodic or cyclical relationships in the data. This research can solve the problem of medium-sized reservoirs that lack sufficient data and modern equipment by enhancing model capabilities even under challenging conditions. RF, in particular, stands out as the most reliable and robust model under these conditions, with cyclical encoding allowing it to achieve its optimal performance. This approach demonstrates that advanced feature engineering can help overcome the limitations associated with inadequate data and outdated infrastructure.

As shown in Table IV, the analysis of feature importance across various models highlights cumulative_outflow and cumulative_inflow_volume as the most influential predictors. In the RF model, Month_Cos (0.3547), Year (0.1451), and cumulative_outflow (0.1283) emerge as the most significant features, reflecting the model's sensitivity to temporal dynamics and aggregated flow data. Additional features, such as Cumulative_ inflow_volume (0.0826) and Wind_current (0.0842), exhibit secondary contributions to prediction accuracy.

TABLE III. PERFORMANCE METRICS OF FOUR MODELS BEFORE USING CYCLICAL ENCODING

Model	Validation set				Test Set			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
RF	182.3616	13.5041	9.4485	0.6485	180.0700	13.4190	9.3966	0.6585
SVM	283.7057	16.8436	12.8562	0.4530	292.9606	17.1161	13.0253	0.4442
GBR with ANN	74.2951	8.6195	6.2368	0.8565	71.5320	8.4577	0.8646	5.8893
LR	179.8926	13.4124	10.4858	0.6525	182.1113	13.4949	10.2469	0.6553

TABLE IV. PERFORMANCE METRICS OF FOUR MODELS AFTER USING CYCLICAL ENCODING

Model	Validation set				Test Set			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
RF	40.1986	6.340	4.3426	0.9223	30.2653	5.501	3.9089	0.9427
SVM	91.660	9.5740	6.6282	0.8229	73.8347	8.5927	6.1695	0.8602
GBR with ANN	84.5176	9.1933	6.0229	0.9031	84.5197	9.1935	6.5001	0.9115
LR	203.0373	14.2491	11.1953	0.6078	201.9911	14.2124	11.1207	0.6177

TABLE V. CONSOLIDATED FEATURE IMPORTANCE COMPARISON

Feature	RF	SVM	ANN (MLPRegressor)	LR
cumulative_outflow	0.128322	0.246976	48.935039	28.235889
cumulative_inflow_volume	0.082626	0.142019	19.130889	33.068566
Month_Cos	0.354743	1.287142	3.643861	1.004209
Year	0.145063	0.552326	4.084095	0.505373
Month_Sin	0.077366	0.302196	2.379926	0.014638
Rain	0.025426	0.019482	0.839323	0.000035
Evaporation	0.049374	0.035942	0.81563	0.000804
Day	0.033126	0.064492	0.083455	-0.001998
wind_current	0.08423	0.039846	0.013068	0.002006
DayOfWeek_Sin	0.009986	0.000451	0.009992	-0.00122
DayOfWeek_Cos	0.007097	-0.001734	0.003729	0.000091
IsWeekend	0.002641	0.001198	0.008334	-0.001511

The ANN model assigns exceptional importance to cumulative_outflow (48.9350) and cumulative_inflow_ volume (19.1309), emphasizing the neural network's capacity to capture complex nonlinear relationships. Features such as Year (4.0841) and Month_Cos (3.6439) provide moderate contributions, whereas features such as IsWeekend (0.0083) and DayOfWeek_Cos (0.0037) have minimal impact. In contrast, the SVM model prioritizes Month_Cos (1.2871), Year (0.5523), and Cumulative_outflow (0.2470), maintaining a balanced emphasis between temporal and cumulative predictors.

LR heavily weights Cumulative_inflow_volume (33.0686) and Cumulative_outflow (28.2359) with moderate contributions from Month_Cos (1.0042) and Year (0.5054). Features such as IsWeekend (-0.0015) and Day (-0.0020) exhibit negative importance, suggesting that their inclusion could detract from model accuracy. Across all models, features such as DayOfWeek_Sin, DayOfWeek_Cos, and IsWeekend generally display negligible or negative importance, indicating their limited relevance for predictive accuracy. This quantitative comparison underscores the dominant role of cumulative metrics in enhancing model performance and highlights the varying priorities of different algorithms.

Both cumulative_outflow and cumulative_inflow_ volumes encapsulate the net effect of hydrological processes over time, providing a comprehensive picture of the system's water balance and effectively capturing temporal dependencies inherent in water systems. The complex relationship between cumulative volumes and water levels, shaped by factors such as evaporation rates, soil absorption, and human interventions, is adeptly modeled by ANNs, which excel at capturing such nonlinearities, making these features highly influential. Models such as RF and SVM prioritize features that effectively split the data. Thus, cumulative metrics often provide significant differences due to their strong correlation with the target variable. LR highlights the strong linear associations of cumulative metrics with the target, leading to high weights, while the ability of ANNs to model complex interactions underscores the critical role these cumulative metrics play in predictive capabilities.

IV. CONCLUSION

This research addresses the challenges faced by medium-sized reservoirs that lack modern equipment by enhancing model capabilities. This study highlights the significant impact of cyclical encoding as a feature-engineering technique in improving the performance of machine learning models, particularly nonlinear algorithms such as RF, SVM, and GBR. The application of cyclical encoding substantially enhanced the predictive power of the RF model, with R^2 values increasing from 0.6485 (validation) and 0.6585 (test) without encoding to 0.9223 and 0.9427, respectively, alongside notable reductions in error metrics, showcasing its ability to effectively capture cyclical patterns in the data. Similarly, SVM and GBR exhibited marked improvements, while LR demonstrated minimal enhancement due to its inherent linear constraints. Feature importance analysis revealed that cumulative outflow and inflow volumes were consistently the most influential predictors across models, with RF effectively leveraging temporal and cyclical features such as Month_Cos (0.3547) and Year (0.1451), and ANN emphasizing nonlinear relationships with high weights on Cumulative_outflow (48.9350) and Cumulative_inflow_volume (19.1309). Conversely, LR relied heavily on cumulative metrics but struggled to utilize encoded features fully, highlighting its limitations. These findings underscore the critical role of advanced feature-engineering techniques such as cyclical encoding in enhancing the accuracy and robustness of nonlinear models, with RF emerging as the most reliable model under all conditions. Future research should explore the application of cyclical encoding in other domains with periodic data, such as climate modeling or financial time series, and investigate hybrid models that combine linear and nonlinear approaches to achieve superior predictive performance and broader applicability.

ACKNOWLEDGMENT

The author extends his sincere gratitude to the Surin Royal Irrigation Project Office 8, Huai Saneng, and the Royal Irrigation Department for providing essential data for this research. Special thanks to Boonlueo Nabumroong and the Department of Computer Technology, Faculty of Agriculture and Technology, RMUTI Surin Campus, for their invaluable support and resources. He is also grateful to his colleagues and peers for their collaboration and feedback and to his family and friends for their unwavering encouragement. Lastly, he appreciates all those whose work indirectly inspired this study.

REFERENCES

[1] C. Carvalho-Santos, A. T. Monteiro, J. C. Azevedo, J. P. Honrado, and J. P. Nunes, "Climate Change Impacts on Water Resources and Reservoir Management: Uncertainty and Adaptation for a Mountain Catchment in Northeast Portugal," *Water Resources Management*, vol. 31, no. 11, pp. 3355–3370, Sep. 2017, https://doi.org/10.1007/s11269-017-1672-z.

[2] S. M. Vicente-Serrano *et al.*, "Global drought trends and future projections," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 380, no. 2238, Oct. 2022, Art. no. 20210285, https://doi.org/10.1098/rsta.2021.0285.

[3] V. Lai, Y. F. Huang, C. H. Koo, A. N. Ahmed, and A. El-Shafie, "A Review of Reservoir Operation Optimisations: from Traditional Models to Metaheuristic Algorithms," *Archives of Computational Methods in Engineering*, vol. 29, no. 5, pp. 3435–3457, Aug. 2022, https://doi.org/10.1007/s11831-021-09701-8.

[4] M. Acreman, "Managed flood releases from reservoirs," presented at the EGS - AGU - EUG Joint Assembly, Apr. 2003, Art. no. 9487.

[5] R. Obringer and R. Nateghi, "Predicting Urban Reservoir Levels Using Statistical Learning Techniques," *Scientific Reports*, vol. 8, no. 1, Mar. 2018, Art. no. 5164, https://doi.org/10.1038/s41598-018-23509-w.

[6] W. J. Wee, N. B. Zaini, A. N. Ahmed, and A. El-Shafie, "A review of models for water level forecasting based on machine learning," *Earth Science Informatics*, vol. 14, no. 4, pp. 1707–1728, Dec. 2021, https://doi.org/10.1007/s12145-021-00664-9.

[7] J. Payen, J. M. Faurès, and D. Vallée, "Small reservoirs and water storage for smallholder farming--the case for a new approach," *Gates Open Research*, vol. 3, no. 387, Feb. 2019, Art. no. 387, https://doi.org/10.21955/gatesopenres.1115486.1.

[8] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I — A discussion of principles," *Journal of Hydrology*, vol. 10, no. 3, pp. 282–290, Apr. 1970, https://doi.org/10.1016/0022-1694(70)90255-6.

[9] H. R. Maier and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications," *Environmental Modelling & Software*, vol. 15, no. 1, pp. 101–124, Jan. 2000, https://doi.org/10.1016/S1364-8152(99)00007-9.

[10] C. W. Dawson and R. L. Wilby, "Hydrological modelling using artificial neural networks," *Progress in Physical Geography*, vol. 25, no. 1, pp. 80–108, Mar. 2001, https://doi.org/10.1177/030913330102500104.

[11] M. S. Khan and P. Coulibaly, "Application of Support Vector Machine in Lake Water Level Prediction," *Journal of Hydrologic Engineering*, vol. 11, no. 3, pp. 199–205, May 2006, https://doi.org/10.1061/(ASCE)1084-0699(2006)11:3(199).

[12] D. P. Solomatine and A. Ostfeld, "Data-driven modelling: some past experiences and new approaches," *Journal of Hydroinformatics*, vol. 10, no. 1, pp. 3–22, Jan. 2008, https://doi.org/10.2166/hydro.2008.015.

[13] I. S. Msiza, F. V. Nelwamondo, and T. Marwala, "Water Demand Prediction using Artificial Neural Networks and Support Vector Regression," *Journal of Computers*, vol. 3, no. 11, Nov. 2008, https://doi.org/10.4304/jcp.3.11.1-8.

[14] H. Liao and W. Sun, "Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method," *Procedia Environmental Sciences*, vol. 2, pp. 970–979, Jan. 2010, https://doi.org/10.1016/j.proenv.2010.10.109.

[15] O. Kisi, J. Shiri, and B. Nikoofar, "Forecasting daily lake levels using artificial intelligence approaches," *Computers & Geosciences*, vol. 41, pp. 169–180, Apr. 2012, https://doi.org/10.1016/j.cageo.2011.08.027.

[16] N. Seckin, M. Cobaner, R. Yurtal, and T. Haktanir, "Comparison of Artificial Neural Network Methods with L-moments for Estimating Flood Flow at Ungauged Sites: the Case of East Mediterranean River Basin, Turkey," *Water Resources Management*, vol. 27, no. 7, pp. 2103–2124, May 2013, https://doi.org/10.1007/s11269-013-0278-3.

[17] C. C. Nwobi-Okoye and A. C. Igboanugo, "Predicting Water Levels at Kainji Dam Using Artificial Neural Networks," *Nigerian Journal of Technology*, vol. 32, no. 1, pp. 129–136, Apr. 2013.

[18] N. Valizadeh and A. El-Shafie, "Forecasting the Level of Reservoirs Using Multiple Input Fuzzification in ANFIS," *Water Resources Management*, vol. 27, no. 9, pp. 3319–3331, Jul. 2013, https://doi.org/10.1007/s11269-013-0349-5.

[19] A. Altunkaynak, "Predicting Water Level Fluctuations in Lake Michigan-Huron Using Wavelet-Expert System Methods," *Water Resources Management*, vol. 28, no. 8, pp. 2293–2314, Jun. 2014, https://doi.org/10.1007/s11269-014-0616-0.

[20] X. Xu *et al.*, "A real-time probabilistic channel flood-forecasting model based on the Bayesian particle filter approach," *Environmental Modelling & Software*, vol. 88, pp. 151–167, Feb. 2017, https://doi.org/10.1016/j.envsoft.2016.11.010.

[21] A. Ogilvie *et al.*, "Combining Landsat observations with hydrological modelling for improved surface water monitoring of small lakes," *Journal of Hydrology*, vol. 566, pp. 109–121, Nov. 2018, https://doi.org/10.1016/j.jhydrol.2018.08.076.

[22] J. Li *et al.*, "Intelligent identification of effective reservoirs based on the random forest classification model," *Journal of Hydrology*, vol. 591, Dec. 2020, Art. no. 125324, https://doi.org/10.1016/j.jhydrol.2020.125324.

[23] Y. O. Ouma *et al.*, "Dam Water Level Prediction Using Vector AutoRegression, Random Forest Regression and MLP-ANN Models Based on Land-Use and Climate Factors," *Sustainability*, vol. 14, no. 22, Jan. 2022, Art. no. 14934, https://doi.org/10.3390/su142214934.

[24] J. Zhang, Y. Zhu, X. Zhang, M. Ye, and J. Yang, "Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas," *Journal of Hydrology*, vol. 561, pp. 918–929, Jun. 2018, https://doi.org/10.1016/j.jhydrol.2018.04.065.

[25] F. Unes, M. Demirci, B. Tasar, Y. Z. Kaya, and H. Varcin, "Estimating dam reservoir level fluctuations using data-driven techniques," *Polish Journal of Environmental Studies*, vol. 28, no. 5, 2019, https://doi.org/10.15244/pjoes/93923.

[26] V. Kaushik and N. Awasthi, "Simulation of reservoir outflows using regression tree and support vector machine," *AI in Civil Engineering*, vol. 2, no. 1, Apr. 2023, Art. no. 2, https://doi.org/10.1007/s43503-023-00012-4.

[27] M. Cho, C. Kim, K. Jung, and H. Jung, "Water Level Prediction Model Applying a Long Short-Term Memory (LSTM)–Gated Recurrent Unit (GRU) Method for Flood Prediction," *Water*, vol. 14, no. 14, Jan. 2022, Art. no. 2221, https://doi.org/10.3390/w14142221.

[28] L. Liu, C. Mou, and F. Xu, "Improved Wildlife Recognition through Fusing Camera Trap Images and Temporal Metadata," *Diversity*, vol. 16, no. 3, Mar. 2024, Art. no. 139, https://doi.org/10.3390/d16030139.

[29] J. Chen and Z. Yang, "Revolutionizing Time Series Data Preprocessing with a Novel Cycling Layer in Self-Attention Mechanisms," *Applied Sciences*, vol. 14, no. 19, Jan. 2024, Art. no. 8922, https://doi.org/10.3390/app14198922.

[30] R. El Moghrabi, R. Tian, L. Liboni, and M. Capretz, "Data Preprocessing for Machine Learning Modules," in *Undergraduate Student Research Internships Conference*, Aug. 2022.

[31] T. Mahajan, G. Singh, and G. Bruns, "An Experimental Assessment of Treatments for Cyclical Data," presented at the Computer Science Conference for CSU Undergraduates, 2021.

[32] A. H. Blasi, M. A. Abbadi, and R. Al-Huweimel, "Machine Learning Approach for an Automatic Irrigation System in Southern Jordan Valley," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6609–6613, Feb. 2021, https://doi.org/10.48084/etasr.3944.

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[34] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.