IJASST
International Journal of
Applied Sciences and Smart Technologies

Volume 07
Issue 02

p-ISSN 2655-8564, e-ISSN 2685-943, pp. 277-298, 2025

Improving the Accuracy of Prediction of Dissolved Oxygen and Nitrate Level Using LSTM with K-Means Clustering and Spearman Analysis

Ika Arva Arshella[1] *, I Wayan Mustika[2], Prapto Nugroho[2]

[1] *Department of Electrical Engineering, Faculty of Science and Technology, Sanata Dharma University, Yogyakarta, Indonesia*

[2] *Department of Electrical Engineering and Information Technology Gadjah Mada University Yogyakarta, Indonesia*

**Corresponding Author: ika_arshella@usd.ac.id*

(Received 15-05-2025; Revised 06-07-2025; Accepted 08-07-2025)

Abstract

This study discusses how to prepare data properly before entering the learning process for prediction using Deep Learning (DL). Long Short-Term Memory (LSTM) is one of the DL methods that is often used for prediction because of its superiority in maintaining long-term information. Although LSTM has proven effective, there are issues related to low-quality data that can reduce prediction accuracy. This problem is important to discuss because accuracy is important in predicting a value while field conditions can reduce the quality of the data taken. Data merging based on the relationship of each data collection location using the Spearman analysis and the K-Means clustering method is used to improve data quality. The results of the study show that improving data quality by merging data using K-Means has been successfully applied to various dataset conditions. In this study, we used two types of datasets related to river water quality, namely Dissolved Oxygen (DO) concentration and Nitrate levels for our simulation. The first data set produced DO predictions for eight locations with an average R^2 = 0.9998, MAE = 0.0007, MSE = $1{,}13\times 10^{-6}$. The second data set produced nitrate predictions for ten locations with an average R^2 = 0.7337, MAE = 0.0111, MSE = 0,00029.

Keywords: Data Merging, K-Means, Long Short-Term Memory, Prediction of Dissolved Oxygen and Nitrate Levels, Spearman

1 Introduction

The pattern of an event can be analyzed and used to predict future events. The pattern is obtained by monitoring of several parameters periodically. This monitoring produces a series of temporal data which is often referred as time series data [1]. In this study, data on river content were used, as changes in water quality can be detrimental to the surrounding ecosystem. Therefore, it is necessary to monitor water quality parameters periodically as an effort to maintain water quality control [2], [3].

The amount of monitoring data that is carried out continuously demands a computational approach that is able to handle the complexity of temporal patterns. Many Deep Learning (DL) models can be used to handle this [1]. Examples of DL computational methods are Convolution Neural Networks (CNN), Deep Belief Networks (DBN), Recurrent Neural Networks (RNN), Auto Encoder (AE), Autoregressive Integrated Moving Average (ARIMA). Long Short-Term Memory (LSTM) and many more. In this study, the LSTM method is used because it has advantages in its architecture that can learn long-term data without losing information. These advantages solve the problem of long-term dependency experienced by RNN [1], [2], [3], [4], [5], [6].

Even though the best computational model has been selected, accuracy still needs to be considered to measure the success of the model. Unnatural changes during data collection can occur and lead to reduced data quality and model performance [2]. Data with low quality will have a negative impact on the basis of decision making, analysis processes and/or predictions of future events [3]. One example of a condition that can negatively impact a model is overfitting. Overfitting is a condition where the model tends to learn the details of values and noise at the training stage, making it difficult to generalize to validation and testing stages [7].

Previous researchers have applied several methods to handle low-quality data. In Rangenatan's research [8], the quality and accuracy of predictions can be significantly improved by performing a good and proper data preprocessing stage. The preprocessing stage consists of data cleaning (outliers, noise, missing values), data integration (correlation analysis, identification), and data reduction (grouping, feature selection).

In the data preprocessing stage, researchers adopted several techniques that have been proven effective in several literatures. Outliers were identified and removed using the Interquartile Range (IQR) method [2], [6], [9]. Linear Interpolation (LI) was applied to handle missing or empty data [2], [5], [6]. The Moving Average (MA) method was used to reduce noise and smooth the data [5], [6]. The selection of this method was based on its application in the literature before the model was developed.

Model development was done by adding the amount of data based on clustering the two most influential attributes in the dataset for each location. The K-Means method is widely used for clustering [10]. Wulandari, et al. [11], used the K-Means method to group and evaluate student performance so that the learning process runs smoothly. Wu et al. [12] and Pangestu et al. [13] combined the use of K-Means and ARIMA. The use of K-Means in [12] to evaluate the level of total phosphorus match with other features. The results were able to reduce the Mean Average Error (MAE) by 44.59% and the Mean Squared Error (MSE) by 56.82% in the prediction process. Chormunge et al. [14] and Gao et al. [15] also use K-Means for optimizing data input by clustering data based on compatibility between the features and produces 90% accuracy. Although there was a decrease in errors, researchers emphasized that there were still obstacles due to low data quality.

The following research is close to the proposed research because it has the same approach. Wei et al. [4] combined water pore pressure data from several river depth points using Partitional Clustering Algorithm (PCA) and predicted using LSTM. The R-Squared (R^2) value increased by 12% compared to predictions without data merging. The use of larger data with additional features is recommended to improve model performance. Arshella et al. [6] used multidimensional input on LSTM in predicting Dissolved Oxygen (DO). This study used the same dataset as the proposed study. The selection of this method can increase the accuracy of water quality prediction for one week up to an R^2 value of 0.999 compared to one-dimensional input. Unfortunately, in the process, it experienced some overfitting and underfitting because the input data still had low quality.

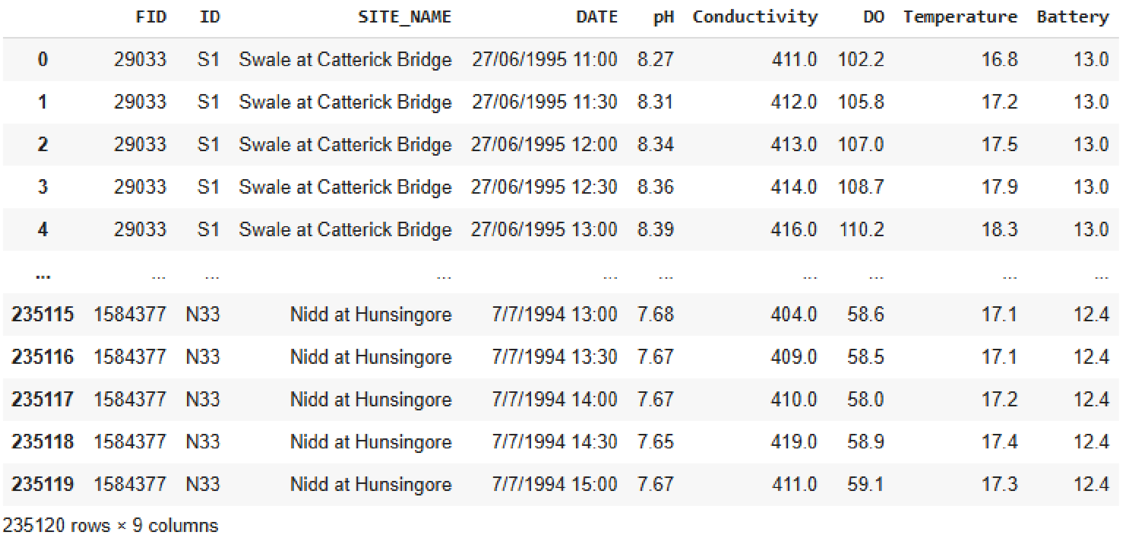
Based on previous studies, data quality has been shown to have a significant impact on the prediction process. The addition of data with similar characteristics is expected to improve data quality with the result prediction accuracy can also increase. This study proposes clustering training data based on the location of water quality data collection using the K-Means clustering method using LSTM. The selection of attributes to use in the K-Means model is determined through the correlation of many attributes to the target attribute of prediction using Spearman method. The model is evaluated using MSE, MAE and R^2.

2 Material and Methods

This study used two water quality datasets obtained from the Environmental Information Data Center, namely the DO level dataset [16] and the Nutrient dataset contained in water [17]. Both datasets are part of the Land Ocean Interaction Study (LOIS) project. The DO dataset used is similar to the dataset used in study [6]. Table. 1 is a view of dataset one. The dataset contains FID as the code of the river location where the data was taken. DATE as the time when the data was taken. Conductivity, pH, Temperature as attributes and DO as the target of the prediction. Data collection was carried out from 1994 to 1997 in locations scattered across the rivers; Swale at Catterick Bridge, Derwent at Bubwith, Aire at Beal Bridge, Trent at Cromwell Lock, Calder at Methley Bridge, Swale at Crakehill, Aire above Thwaite Mill Weir, Aire at Fleet Weir, Ouse at Skelton, Nidd at Hunsingore.

The second dataset is taken from the same website, but contains different water parameters as shown in Table 2. FID is the code of the river location where the data was taken. Date is the time when the data was taken. The amount of data owned in the second dataset only has a small dataset for each location. Datasets for major ions and nutrients in river were collected during the period 1993 to 1997. Locations scattered across the rivers; Swale at Catterick Bridge, Swale at Thornton Manor, Nidd at Skip Bridge, Wharfe at Tadcaster, Ouse at Clifton Bridge, Derwent at Bubwith, Aire at Beal Bridge, Don at Sprotborough Bridge, Trent at Cromwell Lock, Calder at Methley Bridge.

Table 1. DO level dataset



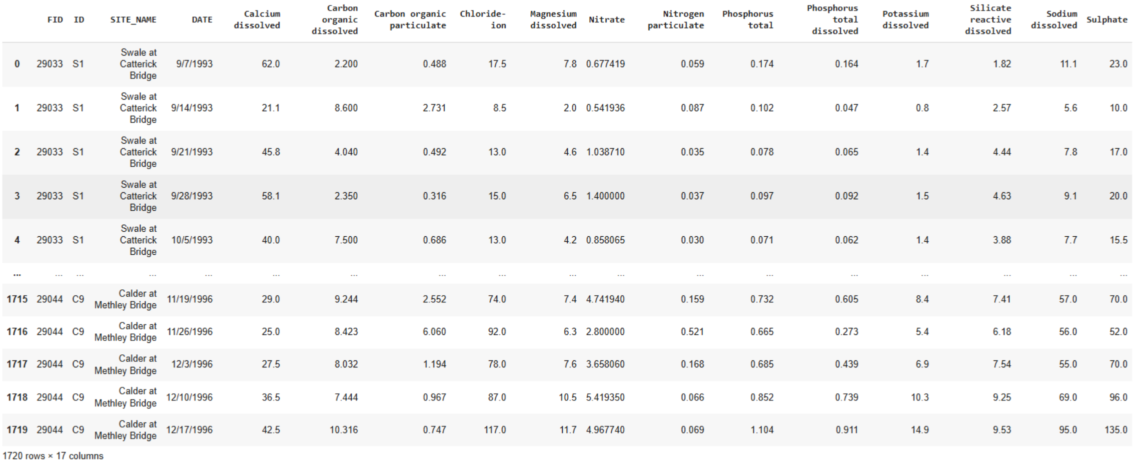
	FID	ID	SITE_NAME	DATE	pH	Conductivity	DO	Temperature	Battery
0	29033	S1	Swale at Catterick Bridge	27/06/1995 11:00	8.27	411.0	102.2	16.8	13.0
1	29033	S1	Swale at Catterick Bridge	27/06/1995 11:30	8.31	412.0	105.8	17.2	13.0
2	29033	S1	Swale at Catterick Bridge	27/06/1995 12:00	8.34	413.0	107.0	17.5	13.0
3	29033	S1	Swale at Catterick Bridge	27/06/1995 12:30	8.36	414.0	108.7	17.9	13.0
4	29033	S1	Swale at Catterick Bridge	27/06/1995 13:00	8.39	416.0	110.2	18.3	13.0
...	...	...	...	...	...	...	...	...	...
235115	1584377	N33	Nidd at Hunsingore	7/7/1994 13:00	7.68	404.0	58.6	17.1	12.4
235116	1584377	N33	Nidd at Hunsingore	7/7/1994 13:30	7.67	409.0	58.5	17.1	12.4
235117	1584377	N33	Nidd at Hunsingore	7/7/1994 14:00	7.67	410.0	58.0	17.2	12.4
235118	1584377	N33	Nidd at Hunsingore	7/7/1994 14:30	7.65	419.0	58.9	17.4	12.4
235119	1584377	N33	Nidd at Hunsingore	7/7/1994 15:00	7.67	411.0	59.1	17.3	12.4

235120 rows × 9 columns

The ions and nutrients measured include Ammonia, Bromide-ion, Calcium Dissolved, Carbon Organic Dissolved, Carbon Organic Particulate. Chloride-ion, Magnesium Dissolved, Nitrate, Nitrite, Nitrogen Particulate, Phosphorus Soluble Reactive, Phosphorus Total, Phosphorus Total Dissolved, Potassium Dissolved, Silicate Reactive Dissolved, Sodium Dissolved, Sulphate. Sampling was carried out regularly every week.

The main flowchart in this study can be seen in Fig. 3. The core of the prediction process begins with loading the water quality dataset into the system, then the data is prepared through a preprocessing stage. At this stage, the data goes through the process of removing outliers and ensuring that there are no missing or empty values so the data used for training will have better quality. When the data is ready, the next step is to predict water quality for the next one-month period using the LSTM algorithm. The prediction results are then analyzed as part of the model evaluation process.

Table 2. Nitrate level dataset



	FID	ID	SITE_NAME	DATE	Calcium dissolved	Carbon organic dissolved	Carbon organic particulate	Chloride-ion	Magnesium dissolved	Nitrate	Nitrogen particulate	Phosphorus total	Phosphorus total dissolved	Potassium dissolved	Silicate reactive dissolved	Sodium dissolved	Sulphate
0	29033	S1	Swale at Catterick Bridge	9/7/1993	62.0	2.200	0.488	17.5	7.8	0.677419	0.059	0.174	0.164	1.7	1.82	11.1	23.0
1	29033	S1	Swale at Catterick Bridge	9/14/1993	21.1	8.600	2.731	8.5	2.0	0.541936	0.087	0.102	0.047	0.8	2.57	5.6	10.0
2	29033	S1	Swale at Catterick Bridge	9/21/1993	45.8	4.040	0.492	13.0	4.6	1.038710	0.035	0.078	0.065	1.4	4.44	7.8	17.0
3	29033	S1	Swale at Catterick Bridge	9/28/1993	58.1	2.350	0.316	15.0	6.5	1.400000	0.037	0.097	0.092	1.5	4.63	9.1	20.0
4	29033	S1	Swale at Catterick Bridge	10/5/1993	40.0	7.500	0.686	13.0	4.2	0.858065	0.030	0.071	0.062	1.4	3.88	7.7	15.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1715	29044	C9	Calder at Methley Bridge	11/19/1996	29.0	9.244	2.552	74.0	7.4	4.741940	0.159	0.732	0.605	8.4	7.41	57.0	70.0
1716	29044	C9	Calder at Methley Bridge	11/26/1996	25.0	8.423	6.060	92.0	6.3	2.800000	0.521	0.665	0.273	5.4	6.18	56.0	52.0
1717	29044	C9	Calder at Methley Bridge	12/3/1996	27.5	8.032	1.194	78.0	7.6	3.658060	0.168	0.685	0.439	6.9	7.54	55.0	70.0
1718	29044	C9	Calder at Methley Bridge	12/10/1996	36.5	7.444	0.967	87.0	10.5	5.419350	0.066	0.852	0.739	10.3	9.25	69.0	96.0
1719	29044	C9	Calder at Methley Bridge	12/17/1996	42.5	10.316	0.747	117.0	11.7	4.967740	0.069	1.104	0.911	14.9	9.53	95.0	135.0

1720 rows × 17 columns

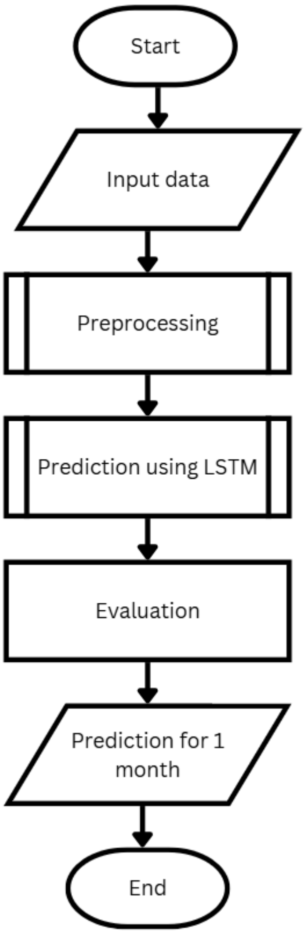


Figure 1. Main flowchart of prediction process

Data Preprocessing

As shown in Figure 1, the first step is reading the input data. The next step is preprocessing, a stage carried out to enhance data quality by removing outliers and verifying the completeness of the dataset. Interquartile Range (IQR) is a method that can be used to identify and remove outliers. By using IQR, data can be selected based on the data area between the upper quartile, which is 75% of the data (Q_3) and the lower quartile, which is 25% of the data (Q_1) as can be seen in Equation (1). Data that is considered to be outliers is data whose value is less than the minimum limit and greater than the maximum limit. Determining the minimum limit using Equation (2) and the maximum limit using Equation (3):

$$\text{IQR} = Q_3 - Q_1, \tag{1}$$

$$\text{Minimum } = (Q_1 - 1.5 * \text{IQR}), \tag{2}$$

$$\text{Maximum } = (Q_3 + 1.5 * \text{IQR}). \tag{3}$$

Where:

IQR : Interquartile Range

Q3 : Upper quartile (75% data)

Q1 : Lower quartile (25% data)

After outliers are removed, the time series data needs to be completed so that there are no null/missing data. Linear Interpolation (LI) is a method used to estimate the value of a point between known data points using a straight line, also known as a linear polynomial. The LI method is widely used because it is simple. LI is used to fill in the gaps experienced by data due to loss and non-empty timing. The new value (y) that is in the middle of the data after (x_1 y_1) and the data before (x_0, y_0) can be calculated based on Equation 4 below:

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}. \tag{4}$$

Where:

x : A point on the x axis is known, its y value is unknown

y	: The point being searched for, using the Equation (4)
(x_1, y_1)	: coordinates new data
(x_0, y_0)	: coordinates new data

From data that has been removed from outliers and filled in, the data is complete according to timing. The data is smoothed or reduced noise using a Moving Average (MA) so the pattern of the data can be displayed clearly and easier to analyze. Moving Averages, also known as running mean or rolling averages, is a special type of filtering method used to transform one time series into another time series. The moving average value for the next period Y_{t+1} is calculated by summing the data values (Y) in a window size (m) and dividing it by the window size value, as shown in the Equation 5;

$$Y_{t+1} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-m+1}}{m}. \tag{5}$$

Where:

Y_{t+1}	: Moving Average value for the next period $(t + 1)$
Y_t	: Data on current time period (t)
Y_{t-1}	: Data before the current time period $(t - 1)$
Y_{t-m+1}	: Data at time period $(t - m + 1)$, back by the window size from the current time period data.
m	: window size, amount of data to calculate the average

K-Means Clustering

The K-Means algorithm uses the proximity measure function to place each object in the cluster that is most similar to it. When updating the center of each cluster (centroid), the distance calculation is performed again and updates the centroid again. This iteration is carried out until the proximity measure function converges; in each cluster the objects no longer change. Iteration is used to divide data objects into several different clusters, so that the similarity between objects in one cluster becomes large, while the similarity between objects becomes small [15]. Calculating the distance between the centroid point and each object point, also called Euclidean distance (D_e), is as in the Equation 6:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}. \tag{6}$$

Where:

D_e	: *Euclidean distance*
(x_i, y_i)	: Object coordinate points
(s_i, t_i)	: Centroid coordinate points

Correlation Analysis

In the K-Means clustering process, distance calculations typically utilize two features. However, the dataset employed in this study contains more than two features, necessitating feature selection. The Spearman method is chosen for feature selection because it is widely used for correlation analysis between features. By using ranking, this method measures the tight relationship between 2 variables. In this case the relationship between the prediction target and other features is calculated using Equation 7:

$$\text{Correlation} = 1 - \frac{6 \times \sum(x-y)^2}{[n \times (n^2 - 1)]}. \tag{7}$$

The results of the Spearman correlation calculation have a value range of -1 to 1. If the result is close to 1, then the two variables have a positive correlation, while the result is close to -1, then the two variables have a negative correlation. If the correlation value is close to 0, then there is no correlation between the two variables [18].

Long Short-Term Memory

Backpropagation errors can lead to signal explosion or vanishing gradients, causing instability during the learning process. To address this issue, Long Short-Term Memory (LSTM) networks are used, as they effectively manage error flow through their gated architecture [19]. The LSTM system consists of three inputs: the current input (x_t), the output of the previous unit (h_{t-1}), and the memory of the previous unit (c_{t-1}).

Figure 2 illustrates the LSTM unit and highlights the three gates that play an important role in the memory calculation: the forget gate, the input gate, and the output gate [2]. The following are the equations used in the LSTM model:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \in (0,1)^h \quad (8)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \in (0,1)^h \quad (9)$$

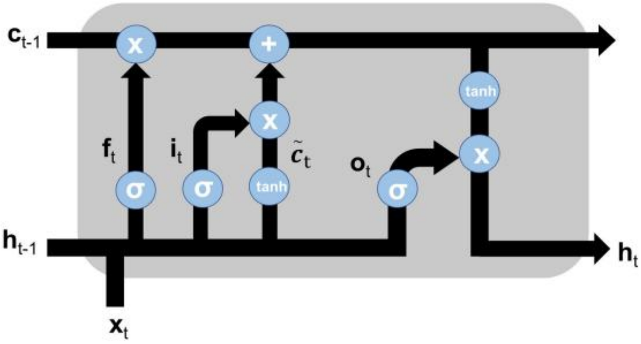


Figure 2. LSTM unit

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \in (-1,1)^h \quad (10)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \in (0,1)^h \quad (11)$$

$$C_t = (f_t \odot C_{t-1}) + \left(i_t \odot \tilde{C}_t\right) \in \mathbb{R}^h \quad (12)$$

$$h_t = o_t \odot \tanh(C_t) \in \mathbb{R}^h \quad (13)$$

Where:

f_t	: Forget gate determinant
i_t	: Input gate determinant
$\tilde{C}_t$	: Determine the memory to be used
o_t	: Output gate determinant
C_t	: New Memory Determinant
h_t	: New Output
W_f, W_i, W_c, W_o	: Weight matrix
b_f, b_i, b_c, b_o	: Bias vector
σ	: Sigmoid function
tanh	: Hypertangen function

Equation (8) is used to calculate the weight vector for the forget gate. It produces an output between 0 and 1 by applying a sigmoid function to the sum of the previous hidden units (h_{t-1}) and the current input (x_t), along with the bias vector for the forget gate (b_f). Similarly, Equation (9) uses the weights and the let vector for the input gate. Equation (10) is used to compute the candidate values to be retained. It produces an output between -1 and 1 by applying the hyperbolic tangent function to the weighted sum of the previous hidden state (h_{t-1}), and the current input (x_t), along with a bias vector for the carrier (b_c). Equation (11) is used to compute the output gate weight vector. It produces the same output as (8) and (9) by using the sigmoid function with output weight matrix (W_o) and output bias vector (b_o).

Equation (12) is used to calculate the memory cell by adding the elementwise multiplication of Equation (8) and the memory of the previous cell (c_{t-1}) with the elementwise multiplication of Equations (9) and (10). Finally, Equation (13) is used to determine the output to be forwarded to the next LSTM cell by multiplying the elementwise multiplication of Equation (11) with the hyperbolic tangent of Equation (12).

Evaluation

A comprehensive evaluation is required to assess the performance of the proposed method. The following evaluation metrics used are mentioned below [20]:

a. Mean Average Error (MAE)

MAE is used in regression analysis to measure the average absolute difference between predicted values (X) and actual values (Y) as in Equation (14):

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i|. \tag{14}$$

Where:

X_i : The i-th predicted value

Y_i : The i-th actual value

m : total data amount

MAE values range from 0 to 1. The closer to 0 the better the results, illustrating that the deviation between prediction and actual is small.

b. Mean Square Error (MSE)

MSE is used in regression analysis to measure the average squared difference between predicted values (X) and actual values (Y) as in Equation (15):

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2. \tag{15}$$

MSE values range from 0 to 1. The closer to 0 the better the results, illustrating that the deviation between prediction and actual is small.

c. R-Squared (R^2)

R^2 is a statistical measure obtained by dividing the mean of the squared differences between the predicted and actual values (MSE) by the total variance of the dependent variable as shown in Equation (16):

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2}. \tag{16}$$

The value of R^2 ranges from -∞ to 1. The closer to 1, the better the result.

3 Results and Discussion

The proposed method was evaluated using two distinct water quality datasets featuring different characteristics and indicators. The analysis is presented in two parts: (1) DO level prediction and (2) nitrate level prediction. Each section systematically compares the performance of the proposed method against baseline approaches.

The result of DO level predictions

The result of DO level predictions from one of the locations can be seen in Fig. 3. The data selected as the visualization of the results comes from the Calder at Methley Bridge with FID 29044. The figure consists of 2 types of graphs, namely loss in training and validation, and a graph of the model's predicted values for one month. Model testing is done with a different preprocessing process. 1) Non-C (nonc), the data used is not merged with data from other locations. 2) Spearman (sp), the data used is merged based on DO correlation between river locations using Spearman method. 3) K-Means (km), the proposed method uses data merging based on clustering attribute data from various locations.

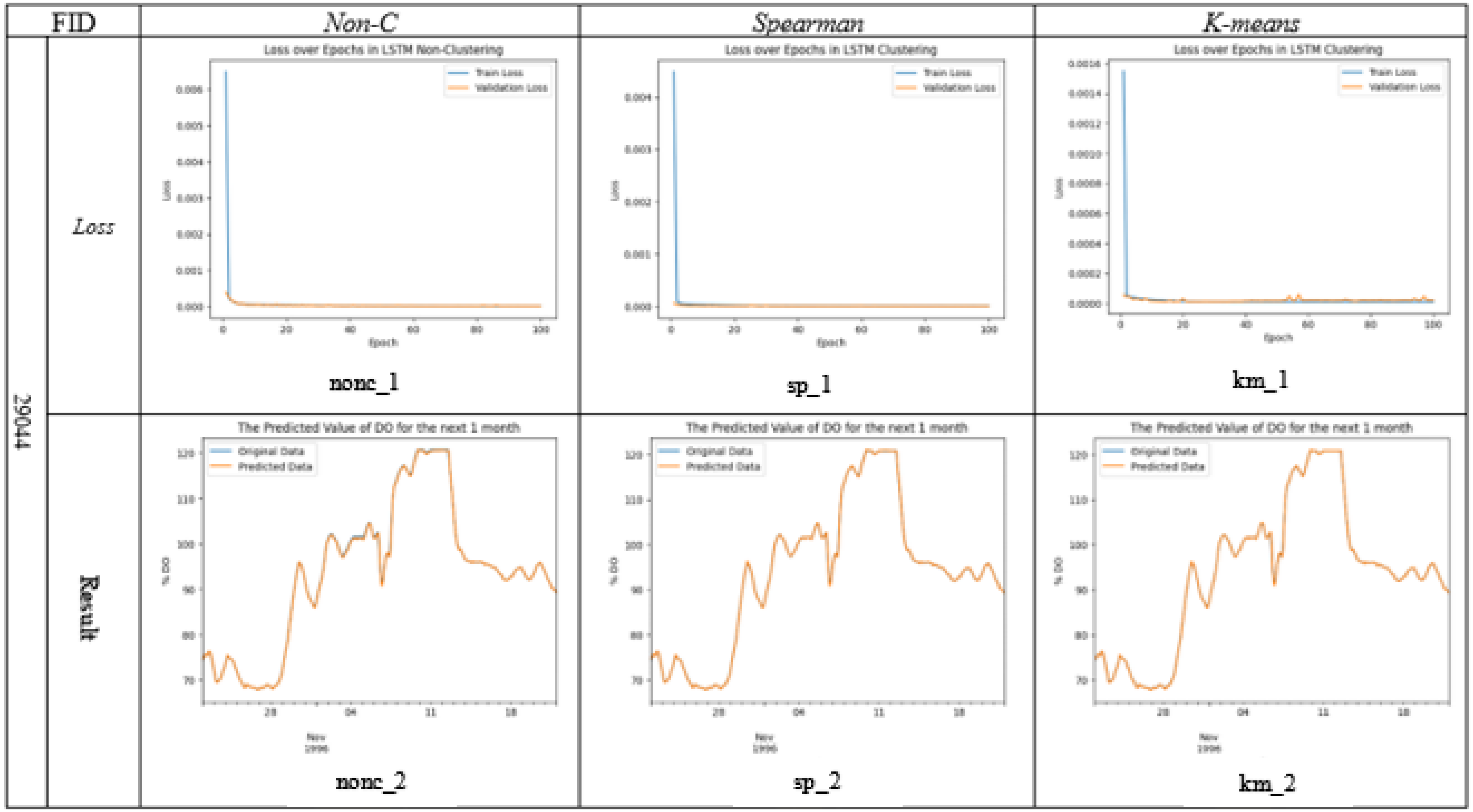


Figure 3. Training and validation loss graphs and DO level prediction results from FID 29044

In Fig. 3(nonc_1, sp_1, km_1), the training and validation loss curves show stable convergence with no sign of overfitting (no significant gap between training and validation loss) or underfitting (loss reaches very low values). All models managed to achieve a loss below 0.001, indicating effective learning. The proposed method (km_1) recorded the lowest loss (less than 0.0002), significantly superior to the nonc and sp baselines. In Fig. 3 (nonc_2, sp_2, km_2), the predicted values of all models almost overlap with the actual values, confirming high accuracy.

Despite the fact that all models appear accurate visually, significant differences are seen in quantitative metrics MAE, MSE and R^2 as can be seen in Table 3. The proposed method reduces MAE for predicting DO in FID 29044 by 61.875% compared to the baseline (nonc), with an absolute value of 0.00061 vs 0.0016. The proposed method reduces MSE by 78.18% compared to the baseline (nonc), with an absolute value of $1{,}2\times 10^{-6}$ vs $5{,}5\times 10^{-6}$. When compared to sp, both (sp and km) have more or less equally good results. This indicates that the application of K-Means for merging data is able to improve model prediction.

The proposed method can produce better values than the comparison method when observed from the average value from various locations. The proposed method produces an average R^2 = 0.9998, MAE = 0.0007, MSE = $1{,}13\times 10^{-6}$. Although statistical significance testing was not performed due to limited data replication, the large

Table 3. Evaluation of the results of the prediction of dissolved oxygen for one month

FID	R²			MAE			MSE		
	nonc	sp	km	nonc	sp	km	nonc	sp	km
29033	0,9996	0,9995	0,9994	0,0016	0,0014	0,0005	$1{,}03\times 10^{-5}$	$6{,}73\times 10^{-6}$	$1{,}13\times 10^{-6}$
29040	0,9999	0,9999	0,9999	0,0012	0,0018	0,0005	$2{,}42\times 10^{-6}$	$3{,}97\times 10^{-6}$	$4{,}10\times 10^{-7}$
29041	0,9997	0,9994	0,9999	0,0014	0,0016	0,0005	$3{,}63\times 10^{-6}$	$2{,}94\times 10^{-6}$	$4{,}87\times 10^{-7}$
29043	0,9999	0,9999	0,9999	0,0014	0,0007	0,0011	$4{,}81\times 10^{-6}$	$1{,}53\times 10^{-6}$	$1{,}94\times 10^{-6}$
29044*	0,9998	0,9999	0,9999	0,0016	0,0007	0,0006	$5{,}5\times 10^{-6}$	$1{,}1\times 10^{-6}$	$1{,}2\times 10^{-6}$
1552046	0,9999	0,9998	0,9998	0,0019	0,0008	0,0007	$6{,}67\times 10^{-6}$	$1{,}17\times 10^{-6}$	$7{,}28\times 10^{-7}$
1552048	0,9991	0,9999	0,9995	0,0024	0,00024	0,0013	$6{,}34\times 10^{-6}$	$1{,}25\times 10^{-7}$	$2{,}39\times 10^{-6}$
1584376	0,9999	0,9998	0,9999	0,0009	0,0013	0,0006	$3{,}18\times 10^{-6}$	$2{,}92\times 10^{-6}$	$7{,}76\times 10^{-7}$
Average	0,9997	0,9997	0,9998	0,0016	0,0011	0,0007	$5{,}36\times 10^{-6}$	$2{,}56\times 10^{-6}$	$1{,}13\times 10^{-6}$

differences and stability of the results support the potential superiority of this method. Further research is needed with more diverse samples, cross-validation, and statistical testing.

The result of nitrate level predictions

The result of nitrate level predictions from one of the locations can be seen in Fig. 4. The data selected as the visualization of the results comes from the Wharfe at Tadcaster with FID 29038. Same as before, the figure consists of 2 types of graphs, namely loss in training and validation, and a graph of model prediction values for one month. Model testing is carried out with different preprocessing processes. 1) Non-C (nonc), 2) Spearman (sp), and 3) K-Means (km). The proposed method uses data merging based on attribute data clustering from various locations.

The gap between training and validation loss in Fig. 4 (nonc_3) suggests overfitting, likely due to the model's high complexity relative to the dataset size. Figure 6 sp_3, km_3, the training and validation loss curves show stable convergence with no sign of overfitting (no significant gap between training and validation loss) or underfitting (loss reaches very low values). All models managed to achieve a loss below 0.005, indicating effective learning. In Fig. 4 (nonc_4, sp_4, km_4), the predicted values of all models almost overlap with the actual values, confirming high accuracy.

Despite the fact that all models appear almost accurate visually, significant differences are seen in quantitative metrics MAE, MSE and R^2 as can be seen in Table 4. The proposed method reduces MAE by 78,5% compared to the baseline (nonc), with an absolute value of 0.016 vs 0.078. The proposed method reduces MSE by 93,7% compared to the baseline (nonc), with an absolute value of 0,0078 vs 0,0005. When compared to sp, both (sp and km) have more or less the same good results but merging using sp correlation tends to be better. This shows that the application of K-Means for nutrient data merging can be used but there is an opportunity to use other methods.

The proposed method can produce better values than the comparison method when observed from the average value from various locations. The proposed method produces an average R^2 = 0.7337, MAE = 0.0111, MSE = 0,00029. This small average is due to the negative R^2 result in one location. Although statistical significance testing was not performed due to limited data replication, the large differences and stability of the

Table 4. Evaluation of the results of the prediction of nitrate for one month

FID	R²			MAE			MSE		
	nonc	sp	km	nonc	sp	km	Nonc	sp	km
29033	-10,4897	-1,07	-1,07	0,0401	0,0056	0,0056	2,40× 10^{-3}	3,99× 10^{-5}	3,99× 10^{-5}
29034	-0,5825	0,8807	0,9232	0,0675	0,0083	0,0057	0,0062	0,0002	1,37× 10^{-4}
29037	0,9417	0,9880	0,9903	0,0359	0,0041	0,0068	0,002	4,47× 10^{-5}	8,64× 10^{-5}
29038*	0,869	0,954	0,958	0,078	0,012	0,016	0,0078	0,0002	0,0005
29039	0,9469	0,9853	0,987	0,029	0,0120	0,0106	0,0012	0,0002	1,74× 10^{-4}
29040	0,7597	0,9717	0,9822	0,0605	0,019	0,0133	0,0046	0,0004	0,0006
29041	0,8325	0,8981	0,9491	0,0178	0,019	0,013	0,0008	0,0005	0,0002
29042	0,7382	0,886	0,8804	0,0646	0,018	0,0166	0,005	0,0007	0,0007
29043	-1,329	0,648	0,6518	0,0455	0,0087	0,007	0,0025	9,28× 10^{-5}	9,18× 10^{-5}
29044	0,648	0,916	0,9051	0,0425	0,012	0,0108	0,003	0,0002	0,0002
Average	-0,666	0,706	0,7337	0,0481	0,0119	0,0111	0,00355	0,00026	0,00029

results support the potential superiority of this method. Further research is needed with more diverse samples, cross-validation, and statistical testing.

In addition to comparing with non-c and sp, comparisons were also made to previous studies. The study conducted by Arshella [6] used the same DO dataset, but the proposed method can overcome the problem of overfitting and can predict for a longer time than previous studies. A similar study was conducted by Wu [12], using K-Means to identify random patterns of water parameters and ARIMA for prediction. Although using a different dataset, the study is still about water quality parameters and using the K-Means method. The method proposed in the study can reduce prediction errors more effectively by overcoming the main limitations in previous studies through preprocessing techniques to improve data quality, as well as combining more advanced learning methods (LSTM) for long-term data analysis. Although successful, this approach still opens up opportunities for future improvements, such as validation on larger or more diverse datasets, exploration of real-time applications, cross-validation, and statistical testing.

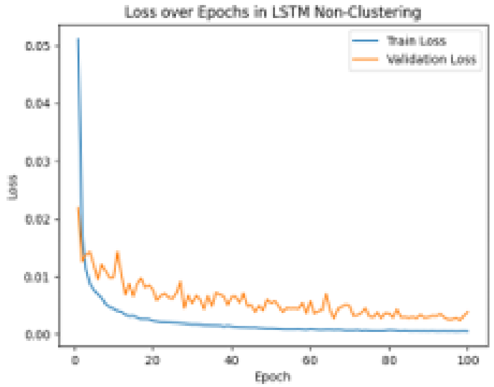
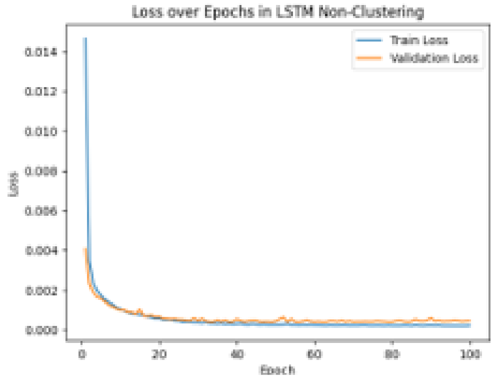
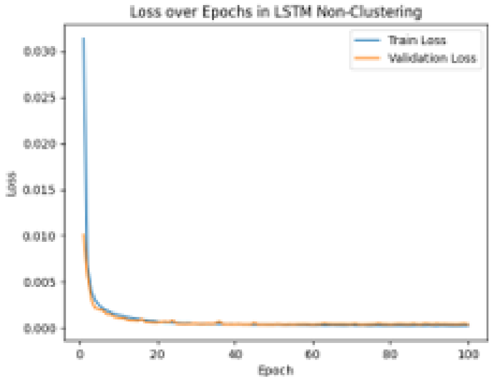
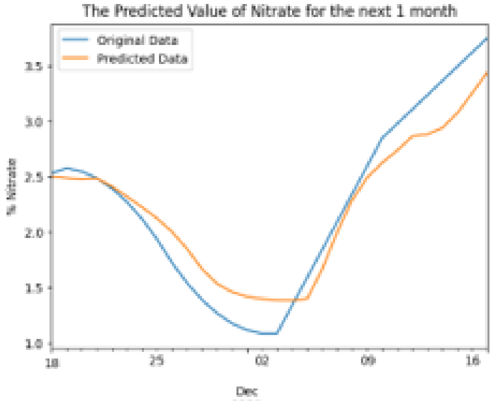
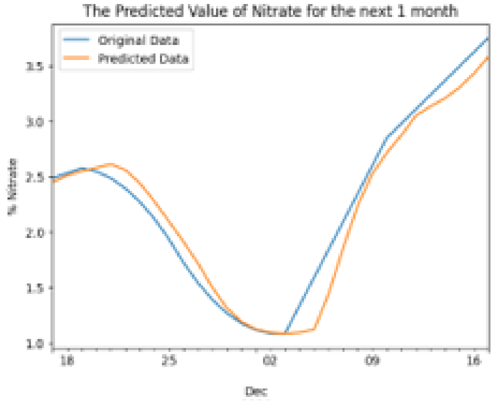
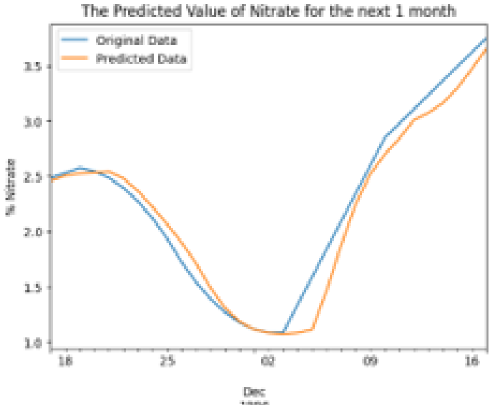
FID		*Non-C*	*Spearman*	*K-means*
29038	*Loss*	 nonc_3	 sp_3	 km_3
	Result	 nonc_4	 sp_4	 km_4

Figure 4. Training and validation loss graphs and nitrate level prediction results from FID 29038

4 Conclusions

The accuracy of water quality prediction can be improved by improving data quality. The application of Spearman method for decide which attributes or parameters need to be used for K-Means, K-Means methods for data merging and LSTM at the prediction stage has a positive impact on increasing accuracy. In this study, we implemented two datasets. The first data set produced DO predictions with an average R^2 = 0.9998, MAE = 0.0007, MSE = $1{,}13\times 10^{-6}$. The second data set produced nitrate level predictions with an average R^2 = 0.7337, MAE = 0.0111, MSE = 0,00029. Data merging based on location can be applied to various dataset conditions; datasets with few features and many data and datasets with few data and many features. The selection of the right grouping method should also be further evaluated to ensure optimal results. These steps will be an important part of the development and assessment of prediction methods, especially for water quality in the future. Other suggestions for improvement are the use of larger or more diverse datasets, real-time implementation, cross validation, and statistical testing.

Acknowledgements

The author would like to thank the supervisor from Gadjah Mada University who has guided this research so that this research can be completed.

References

[1] H. Zhongyang, Z. Jun, L. Henry, F. M. King, and W. Wei, “A review of deep learning models for time series prediction,” *IEEE Sensor Journal*, vol. 21, No. 6, Mar. 2021, doi: 10.1109/JSEN.2019.2923982.

[2] H. Chen *et al.*, “Water quality prediction based on LSTM and attention mechanism: A case study of the Burnett River, Australia,” *Sustainability (Switzerland)*, vol. 14, no. 20, Oct. 2022, doi: 10.3390/su142013231.

[3] A. Docheshmeh Gorgij, G. Askari, A. A. Taghipour, M. Jami, and M. Mirfardi, “Spatiotemporal forecasting of the groundwater quality for irrigation purposes, using deep learning Method: Long short-term memory (LSTM),” *Agric Water Manag*, vol. 277, Mar. 2023, doi: 10.1016/j.agwat.2022.108088.

[4] C. W. W. Ng, M. Usman, and H. Guo, "Spatiotemporal pore-water pressure prediction using multi-input long short-term memory," *Eng Geol*, vol. 322, Sep. 2023, doi: 10.1016/j.enggeo.2023.107194.

[5] Z. Hu *et al.*, "A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture," *Sensors (Switzerland)*, vol. 19, no. 6, Mar. 2019, doi: 10.3390/s19061420.

[6] Arshella. Ika Arva, I. W. Mustika, and P. Nugroho, "Water quality prediction based on machine learning using multidimension input LSTM," IEEE, Aug. 2023. doi: 10.1109/ICITACEE58587.2023.10276970.

[7] M. Del Giudice, "The prediction-explanation fallacy: A pervasive problem in scientific applications of machine learning," *Methodology*, vol. 20, no. 1, pp. 22–46, 2024, doi: 10.5964/meth.11235.

[8] R. G, "A study to find facts behind preprocessing on deep learning algorithms," *Journal of Innovative Image Processing*, vol. 3, no. 1, pp. 66–74, Apr. 2021, doi: 10.36548/jiip.2021.1.006.

[9] D. Dheda, L. Cheng, and A. M. Abu-Mahfouz, "Long short-term memory water quality predictive model discrepancy mitigation through genetic algorithm optimisation and ensemble modeling," *IEEE Access*, vol. 10, pp. 24638–24658, Feb. 2022, doi: 10.1109/ACCESS.2022.3152818.

[10] M. G. H. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," 2007, *IOS Press*. doi: 10.3233/ida-2007-11602.

[11] N. H. Wulandari and V. Purwayoga, "Cluster change analysis to assess the effectiveness of speaking skill techniques using machine learning," *International Journal of Applied Sciences and Smart Technologies*, vol. 7, no. 1, pp. 1–14, 2025, doi: 10.24071/ijasst.v7i1.9667.

[12] J. Wu *et al.*, "Application of time serial model in water quality predicting," *Computers, Materials and Continua*, vol. 74, no. 1, pp. 67–82, 2023, doi: 10.32604/cmc.2023.030703.

[13] P. Pangestu, S. Maarip, Y. N. Addinsyah, and V. Purwayoga, "Clustering and trend analysis of priority commodities in the archipelago capital region (IKN) using a data mining approach," *International Journal of Applied Sciences and Smart Technologies*, vol. 6, no. 1, pp. 169–182, 2024, doi: 10.24071/ijasst.v6i1.7798.

[14] S. Chormunge and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 542–549, Dec. 2018, doi: 10.1016/j.jesit.2017.06.004.

[15] G. Qiang, X. Hong Xia, H. Hong Gui, and G. Min, "Soft sensor method for surface water qualities based on fuzzy neural network," IEEE, Jul. 2019. doi: 10.23919/ChiCC.2019.8866494.

[16] D. Leach, A. Pinder, P. Wass, N. Bachiller-Jareno, I. Tindall, and R. Moore, "Continuous Measurements of Temperature, pH, Conductivity and Dissolved Oxygen in Rivers [LOIS]," NERC Environmental Information Data Centre. Accessed: Aug. 01, 2023. [Online]. Available: https://doi.org/10.5285/b8a985f5-30b5-4234-9a62-03de60bf31f7

[17] D. Leach, M. Neal, N. Bachiller-Jareno, I. Tindall, and R. Moore, "Major ion and nutrient data from rivers [LOIS]," NERC Environmental Information Data Centre. Accessed: Aug. 01, 2023. [Online]. Available: https://doi.org/10.5285/4482fa14-aee2-4c7f-9c62-a08dc9704051

[18] Centre for Innovation in Mathematics Teaching, *Correlation and regression*. University of Plymouth. Accessed: Jun. 27, 2025. [Online]. Available: https://www.cimt.org.uk/projects/mepres/alevel/stats_ch12.pdf

[19] H. Sepp and S. Jurgen, “Long short-term memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[20] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Comput Sci*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

This page intentionally left