Artificial Neural Network Backpropagation for Real-Time Coagulant Dose Prediction in Drinking Water Treatment

Hilmi Putra Pradana[1], *Berliana* Khansa Salsabila[1], *Achmad* Muzakky[1], *Mas* Agus Mardianto[1], and *Ervin* Nurhayati[1*]

[1] Department of Environmental Engineering, Faculty of Civil, Planning, and Geo Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
* Corresponding author: ervin@its.ac.id

Abstract. A drinking water treatment plant is crucial for fulfilling the increasing water demand. As an integrated series, the coagulation unit is the most basic unit for removing particulates and reducing turbidity. However, the time gap to determine an effective coagulant dose was almost 6 h, which cannot accommodate the fluctuations in water inlet quality. A noteworthy method to reduce time is to use artificial neural network backpropagation (ANN-BP) to predict an optimum dose. The dataset consisted of five parameters (pH, temperature, conductivity, color, and turbidity) for a month of primary data sampling and historical jar test data from 2018 to 2022. The F-test result, F-value (6038,779) > F-table (2.21923), showed that one or more parameters had a statistically significant influence on the coagulant dose. Subsequently, a t-test excluded pH and temperature, with p-values lower than 0.05. Empirical models were developed through trial-and-error variations of the input layers (three and five parameters), hidden layers (2-10 nodes), and an output. The models with the lowest MSE and highest R^2 were [5-6-1] (R^2 =- 0.96051; MSE = 0.00179) and [3-4-1] (R^2 = 0.97755; MSE =0.00102). In conclusion, [3-4-1] is recommended because it has the lowest MSE and the highest R^2.

Keyword: ANN-BP, Coagulation, Drinking water, Water treatment

1. Introduction

The necessity of potable water has been rising over the years as domestic and industrial water use increases. Surabaya's water demand in 2039 was projected to be 15,528 L/s from 7,610 L/s in 2018 as the baseline value [1], while Gombi, Nigeria's water demand was forecasted to increase by 54.5% in 2031 compared to 2021 [2]. To meet the increasing demand for water quantity and the need to meet potable water quality standards, non-potable natural water can be treated using a drinking water treatment plant (DWTP), which consists of an integrated series of operations and process treatment units [3]. One of the most basic unit processes commonly and fundamentally used in DWTP is the coagulation unit paired with flocculation to remove particulates and reduce turbidity [4].

As an essential part of the treatment process, coagulation is a destabilization process of suspension or solution by adding a coagulant (a substance to give an effect of destabilization), which induces rapid turbulence to form visible floc or precipitate that will be separated by gravitational force or filtration [5]. The optimum coagulant doses are normally calculated using the jar test method, which requires a time to conduct the test of approximately half an hour to 6 hours from sampling until the outcome is obtained [6–8]. The time gap between them cannot accommodate the variety of water quality parameters that influence the coagulation-flocculation process or, in general, the water treatment process [6]. The negative effects of unfitted doses may result in increased operational costs, inefficient removal, and a risk to consumer safety [6]. Therefore, there is a need to discover a quicker technique to determine optimum coagulant doses.

A noteworthy method that has been gradually explored by many studies involves machine learning or artificial intelligence to quickly predict coagulant doses based on fluctuations in inlet water quality. Various models have been used, such as partial least squares regression (PLS), adaptive neuro-fuzzy inference system (ANFIS), multiple linear regression (MLR), and artificial neural network (ANN) [6–8]. Every study stood up by combining or comparing several methods, by using different input parameters, or various data sizes.

This study aims to determine the effectiveness of coagulant dose prediction based on a back-propagation ANN method from five regular water parameters, where the result can be obtained in near real-time from

the outlet of the prior reactor or the inlet coagulation process, such as pH, temperature, conductivity, color, and turbidity. The key parameters would be extracted F-test and T-test to choose the most influential parameters. A comparison between the full set of five parameters and the statically selected parameters for building the ANN model. The performance of each model was evaluated to determine which model best represented the jar test results based on the coefficient of determination (R^2) and mean square error (MSE). Moreover, this study emphasized the use of a large historical dataset as a secondary source and primary parameter sampling to enhance the confidence and reliability of the model and minimize prediction error.

2. Methodology

2.1 Data sources

For this study, primary and secondary data were used. Primary data were acquired from real-time sampling from Karangpilang and Ngagel DWTP for a month. Samples were collected nine times a day (08.00 AM, 08.30 AM, 09.00 AM, 12.00 AM, 12.30 PM, 13.00 PM, 15.00 PM, 15.30 PM, and 16.00 PM) from the water outlet of the pre-sedimentation tank or the inlet of the coagulation unit. A 10-litre sample volume was taken for each period, then the five field parameters were measured (pH, turbidity, temperature, color, and conductivity). Subsequently, the total sampling volume was divided into 5 container sub-samples and 1 blank container for jar test analysis. Each container consisted of a 1 L sample. 1% liquid aluminium sulfate, $[Al_2(SO_4)_318H_2O]$, was the coagulant used in this study, as this substance is commonly utilised in Indonesia's DWTP, as well as in the location of the study object.

Jar test analysis was performed by applying 5 variations of coagulant doses separately. Mixing was arranged at 100 rpm for a minute as rapid mixing, then 40-60 rpm for 15 minutes as slow mixing phase. Finally, the settling time was set at 15 minutes. Optimum coagulant doses were chosen if the turbidity parameter was < 5 NTU; additionally, after treatment, the parameters were also tested.

Historical data, as a secondary source, were provided by the Surabaya Regional Water Company as the regulator of the primary sampling location. The data ranged from 2018 to 2022, with a total of 10282 data. Each set contained the same parameters that were obtained by primary sampling.

2.2 Determining the significant parameter affecting the dose of coagulant

Five parameters as independent variables and coagulant doses as dependent variable were statistically analyzed with F-test, while the influence of each parameter was also analyzed with pairwise t-test. Significant parameters from the individual t-test will be used as another input layer in the second ANN model, in addition to all the initial parameters in the first ANN model.

2.3 ANN model configuration

In this study, back-propagation ANN (ANN-BP) models were used. Fundamental idea of backpropagation, as primary algorithm to compute gradients efficiently in ANNs, is about updating the weights and biases of the network from the way that decreases the error or cost function by iteration process [9]. ANN-BP, as a supervised learning algorithm, consisted of three layers, namely, an input layer consisting of the values of water quality parameters, a hidden layer consisting of 2 – 10 nodes, as the exact node numbers would be determined from trial-error, and an output layer for coagulant dose values.

The ANN-BP process in this study was implemented using MATLAB. The process involved a training phase (80% of the dataset), followed by a testing phase (20% of the dataset). Each phase the input and output data would be normalized to synthesize ANN-BP model, then the results also should be denormalized. Function for process building ANN architecture in training and testing phase were 'newff' to build feed-forward backpropagation ANN, while simulation in each phase after build ANN model used 'sim' function. In the training phase, the ANN architecture parameters were 0.5 for the learning rate, 1000 for the maximum iterations (epochs), and 0,9 for the momentum parameter. Activation functions were assigned to link the layers: a sigmoid function (logsig) for the input layer and a linear function (pureline) for the output layer. The model's performance was assessed based on its accuracy in predicting the coagulant dose in comparison to the experimental jar test results from the coefficient of determination (R^2) and mean square error (MSE).

3. Result and Discussion

3.1 Significant Parameters Affected Coagulant Doses

The primary and secondary data of five parameters were statistically analysed with F-test to identify the significant effect of multiple parameters in wastewater to coagulant doses. The result shown (Table 1) that F-value (6038.779) was higher than F-table (2.21923) with significancy below α (<0.05) [10]. It can be concluded that there is one or more parameters (pH, color, temperature, conductivity, and turbidity) that have a statistically significant influence on the coagulant dose or the overall combination of all parameters affecting the coagulant dose [10].

Table 1. F-test result value

Model	df	Mean Square	F	Sig
Regression	5	1469945.479	6038.779	.000[b]
Residual	1741	243.418		
Total	1746			

After knowing the F-test result, a t-test for each parameter effect on coagulant doses was conducted. The test was applied to determine the effect of each parameter on the coagulation by comparing the change of parameter and its effect of the coagulant doses.

The statistical outcome (Table 2) revealed that the t-test significance value of temperature (0.658) and pH (0.142) parameters didn't influence coagulant doses because their significance level is higher than α (0,05). pH in this study may not influence the coagulant dose because the narrow value change on the pH range that was between 5.5 – 8.5, as the pH range is still within the optimum pH for alum [6]. This was because the type of coagulant was chosen based on the pH range of the water input of the water treatment plant. A similar effect had occurred in the temperature parameter. The slight range between 28°C - 33°C may not significantly effluence the coagulant dose in the coagulation process. Even though the change in temperature and pH can alter water viscosity and ionic charge, the increase in alum dose for optimal coagulation [11]. The variation of pH and temperature in this study was insufficient to produce a measurable difference. As a result, no statistically significant variation in coagulant doses was found.

In contrast, turbidity, color, and conductivity indicated influence of each parameter for the coagulant dose need, as those showed their significance value are lower than 0.05. The change in turbidity affects the need for coagulant dose adjustment. The higher the turbidity, the more coagulant doses are required. Increasing turbidity escalate the suspended solid that contains ionic strength with same charge, so to attract the whole suspended solid with enough opposite charge is higher comparing with water with low turbidity [12]. Interestingly, water with lower turbidity requires a higher coagulant dose ratio, while an increase in turbidity does not significantly increase the coagulant dose. It because the collision efficiency in high turbidity will increase then improving particle-aggregation caused by neutralization [13]. Color describes the presence of dissolved organic substances that are usually negatively charged, thereby increasing the consumption of coagulants through complex adsorption and bonding mechanisms. Conductivity, as an indicator of ionic strength, affects the thickness of the electrical double layer in the particles and changes the destabilization efficiency of the charge [14]. Considerable variation in these three parameters results in a significant difference in the dosage requirement of coagulants, so it is statistically significant.

Table 2. T-test result value

Model	Standardized Coefficients	t	Sig.
(constant)		-2.576	.010
Turbidity	.343	25.396	.000
Temperature	-.003	-.443	.658
pH	.008	1.467	.142
Color	.441	6.353	.000
Conductivity	.267	4.704	.000

Based on the results of this t-test, turbidity, color, and conductivity can be considered as alternative input layers in the Artificial Neural Network (ANN) model. Using only significant parameters can simplify model architecture, improve training efficiency, and reduce the risk of overfitting without sacrificing the accuracy of coagulant dose predictions.

3.2 Configurations of the ANN model

Two configuration model input layers were used as the main alternatives, as well as the neuron form hidden and output layer. The input layer variations contained five neurons as all the influent water parameters and three neurons, consisting of turbidity, color, and electric conductivity, because of the prior significant parameter test. The model parameters and properties can be seen in the Table 3.

Table 3. The model parameters and properties

Parameter	Value	Description
Input *Layer*	5 neurons	Turbidity, pH, temperature, color, electrical conductivity
	3 neurons	Turbidity, color, electrical conductivity
Hidden Layer	2-10 neurons	Trial and error
Output *Layer*	1 neuron	Optimal dose of coagulants
Learning Rate	0.5	-
Momentum	0.9	-
Epoch	1000	Minimum value
Fungsi Aktivasi	2	Logsig and purelin

The results of the model training and testing for the ANN showed variation in performance between the configuration models because of the variation in neurons in the input and hidden layers. This analysis focused on comparing the performance of all models based on two metrics, the mean squared error (MSE) and the determination coefficient (R^2), to measure the average error rate and the relationship between the input and output. The output MSE and R^2 of model configurations can be seen in Table 4 and illustrated in Figure 1a and 1b.

Table 4. The output MSE and R^2 of model configurations

Input layer	Hidden layer (node)	Output layer	Model	MSE	R^2
5 (pH, temperature, conductivity, color, turbidity)	2	1 (Coagulant dose)	[5,2,1]	0.00251	0.94422
	3		[5,3,1]	0.00820	0.80381
	4		[5,4,1]	0.00601	0.86042
	5		[5,5,1]	0.00266	0.94079
	6		[5,6,1]	0.00179	0.96051
	7		[5,7,1]	0.00870	0.79029
	8		[5,8,1]	0.00287	0.93587
	9		[5,9,1]	0.01096	0.72610
	10		[5,10,1]	0.01631	0.54459
3 (conductivity, color, turbidity)	2		[3,2,1]	0.00153	0.96631
	3		[3,3,1]	0.00152	0.96664
	4		[3,4,1]	0.00102	0.97755
	5		[3,5,1]	0.00537	0.87652
	6		[3,6,1]	0.00347	0.92201

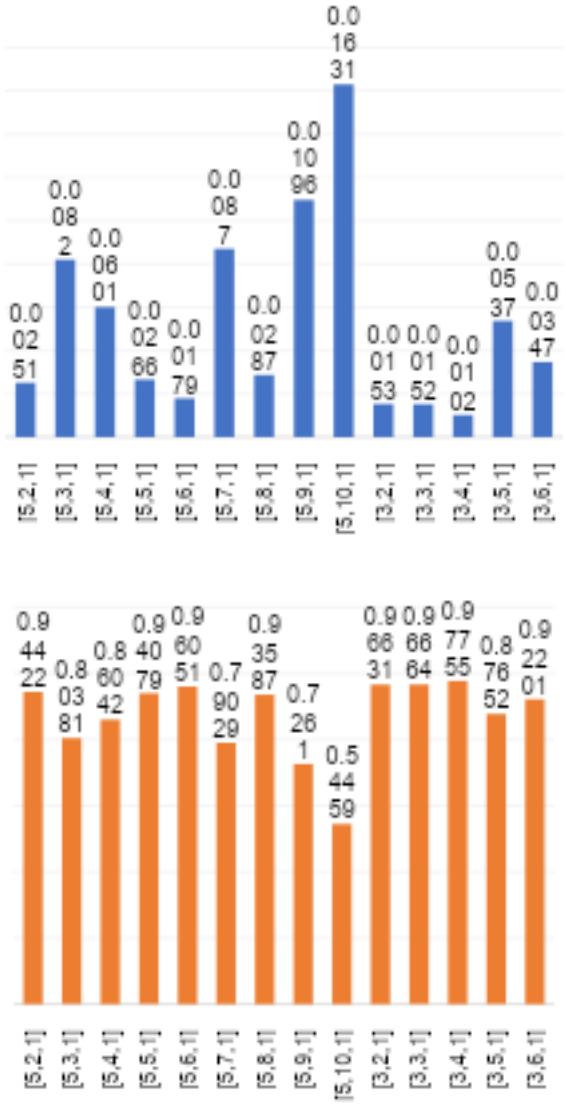

Fig. 1. The fluctuation of MSE (1a) and R^2 (1b) from the model configuration

In the configuration of the input layer with 5 neurons, the model performance shows a fluctuating pattern as the number of neurons in the hidden layer increases. The model with configuration [5,2,1] yielded an MSE of 0.00251 with an R^2 of 0.94422, indicating that the model's ability to explain data variations was quite high. However, when the number of neurons in the hidden layer was increased to 3, 4, and 5, the MSE value amplified at some point, indicating that the increase in model complexity did not necessarily result in increased accuracy and become overfitting because of increasing of hidden layer [15]. Of all the experiments with 5 input neurons, model [5,6,1] showed the best results with an MSE of 0.00179 and an R^2 of 0.96051. A low MSE value indicates a small prediction error rate, whereas a high R^2 value indicates that the model can account for approximately 96% of the variation in the output data. This means that the relationship between water quality parameters (turbidity, pH, temperature, color, and electrical conductivity) and the optimal coagulant dose can be well represented by the model. However, adding more than six neurons significantly reduced the model performance. For example, in model [5,9,1], the MSE value increases to 0.01096 and R^2 decreases to 0.72610, indicating that the model is becoming less stable and is prone to overfitting [15]

Meanwhile, in the input layer configuration with 3 neurons, the overall performance of the model showed better and stable results compared to the 5-neuron input configuration. It also shows that the use of fewer but relevant input parameters can result in more efficient models without sacrificing the prediction accuracy [15]. The best results were achieved by the model [3,4,1], with an MSE of 0.00102 and an R^2 of 0.97755. This value was the highest among all the configurations tested, indicating that the model had excellent predictive capabilities and a low error rate. An R^2 value close to 1 indicates that the model can explain almost the entire variation in the

output data; in other words, the relationship between the three input parameters and the optimal coagulant dose can be very well represented by this model. In addition, these results also indicate that the addition of parameters such as pH and temperature (as in the configuration of 5 input neurons) does not provide a significant increase in accuracy and may even magnify the noise in the data and decrease the stability of the model [15].

3.3 Best ANN model and limitation model application

The Artificial Neural Network (ANN) model is used to predict the need for coagulant doses based on raw water quality parameters. Two model architectures were tested, namely the 5–6–1 model and the 3–4–1 model, which was found to be the best model configuration at the input layer of 5 neurons and 3 neurons. From these variations, both architectures demonstrated optimal performance within their respective structures, and their predictive behavior was further evaluated to determine the most reliable model. ANN-BP architectural for 5-6-1 model configuration can be seen in Fig 2.

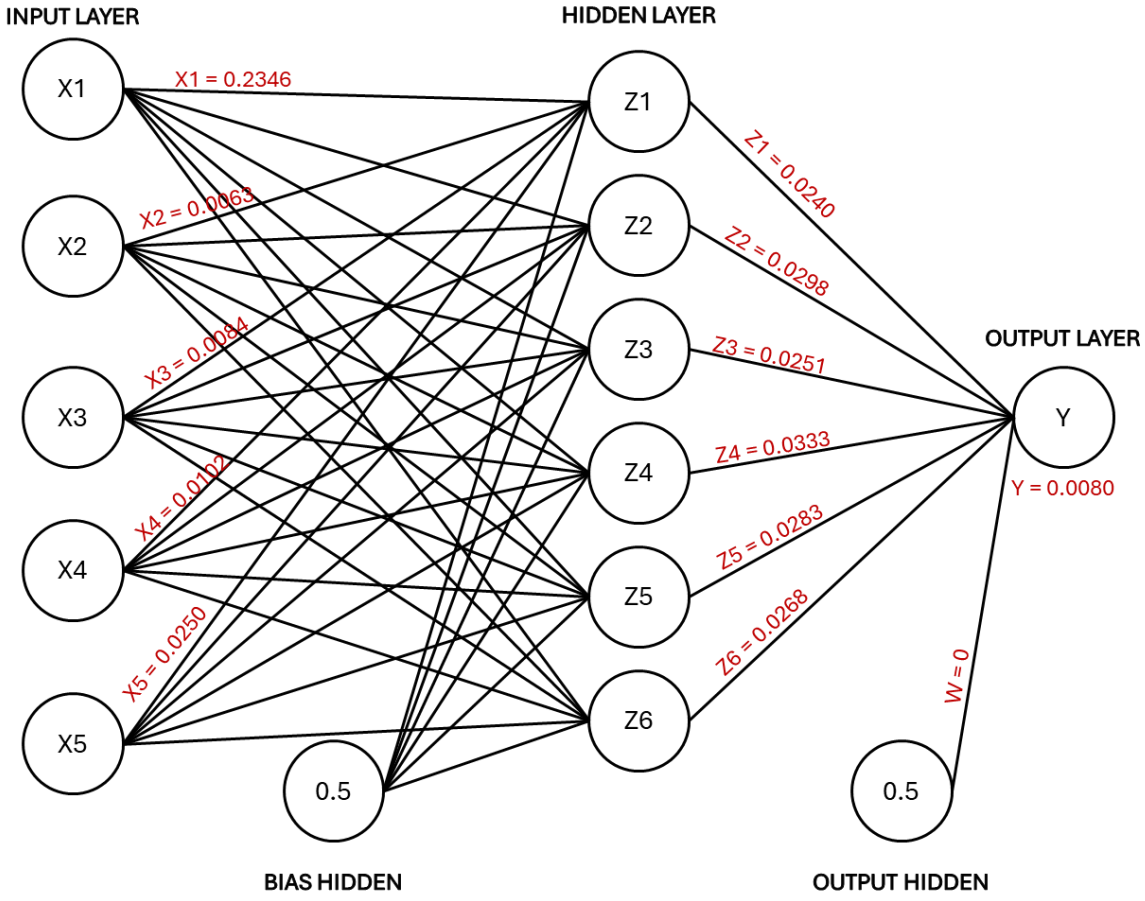

Fig. 2. ANN-BP architectural for 5-6-1 model configuration

For the 5–6–1 architecture, the model with five input neurons generates the normalized predictive equation shown in Equation (1).

coagulant dose = (turbidity x 0.0268) - (temperature x 0.0241) + (pH x 0.0241) + (color x 0.0242)
+ (electrical conductivity x 0.0243) (1)

After denormalization, the equation becomes (Equation 2).

coagulant dose = (turbidity x 0.298) -(temperature x 0.138) + (pH x 0.134) + (color x 0.165)

+ (electrical conductivity x 0.163) (2)

From the equation 1 and 2, turbidity is the variable with the greatest influence, indicating that increasing the turbidity of raw water will significantly increase the need for coagulants. The color and conductivity parameters have a considerable positive influence, indicating their relationship to the concentration of organic matter and the content of dissolved ions that affect the coagulation process. On the contrary, temperature exerts a negative influence, which means that rising temperatures tend to decrease the need for coagulants. This has to do with increasing the kinetic energy of particles at higher temperatures, so that the charge destabilization process can take place more effectively.

On the other hand, the result of ANN-BP architectural for 5-6-1 model configuration can be seen in Fig 3.

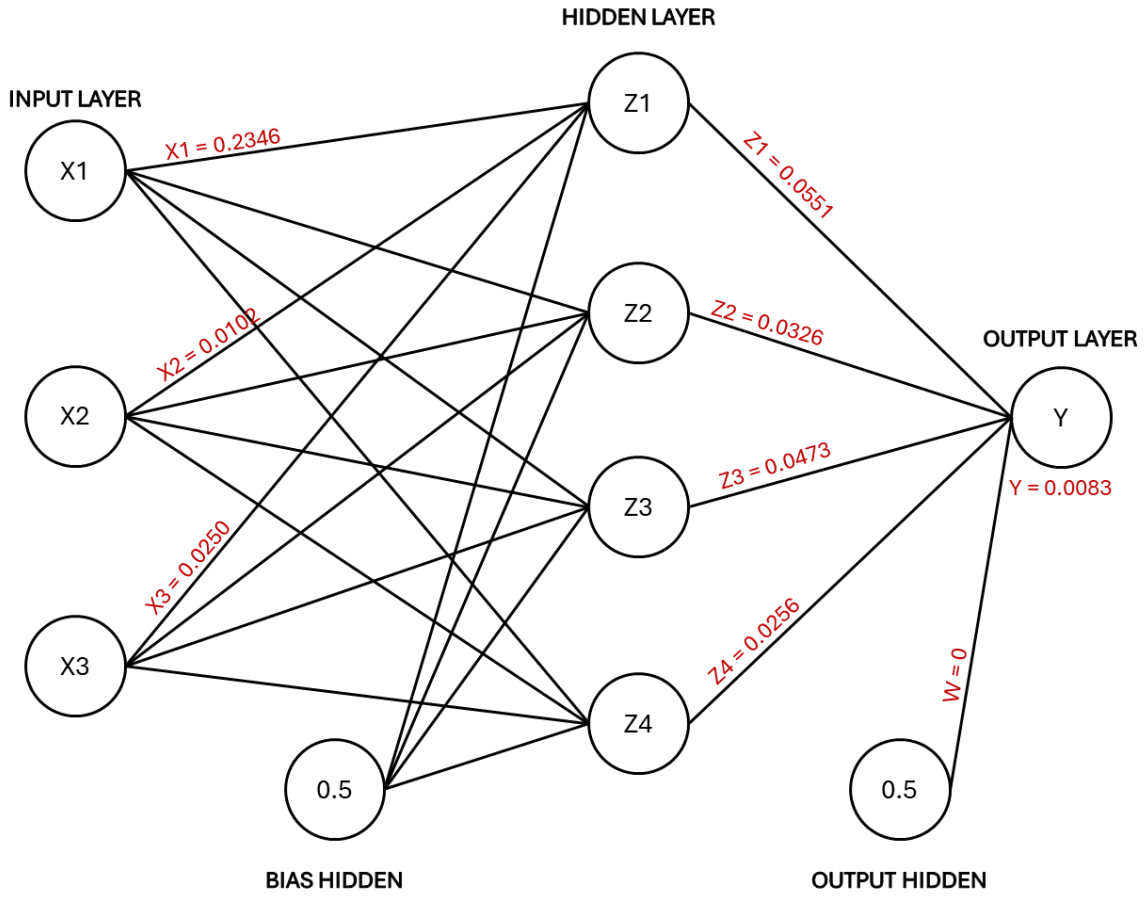

Fig. 3. ANN-BP architectural for 3-4-1 model configuration

The 3–4–1 architecture produces a simplified normalized model shown in Equation (3):

$$coagulant\ dose = (turbidity\ x\ 0.0414) + (color\ x\ 0.0373) + (electrical\ conductivity\ x\ 0.0376) \quad (3)$$

Following denormalization, the model becomes (Equation 4)

$$coagulant\ dose = (turbidity\ x\ 0.306) + (color\ x\ 0.159) + (electrical\ conductivity\ x\ 0.151) \quad (4)$$

Similar to the previous model, turbidity remains the dominant variable, with color and conductivity continuing to serve as meaningful predictors in describing organic content and total dissolved solids that influence coagulation efficiency. This simplified input structure highlights the model's ability to perform effectively even with fewer parameters, focusing on the variables most directly associated with coagulant demand.

In comparison, the [3,4,1] model has the most superior performance with the highest R^2 (0.97755) and the lowest MSE among all models, making it the most efficient and accurate model. While the model [5,6,1] still provides excellent results, it is more complex and more sensitive to data variations. This ANN model is built on data with a specific range, so its use must pay attention to limitations. This is because the model becomes accurate only when the temperature is in the range of 24°C – 33°C according to the air temperature according to BMKG (2022) and pH 5.5 – 8.5 according to the optimal pH of the use of alum-type coagulants. This limitation is important because ANN studies patterns based on available data. When the input is outside of that range, the model no longer describes the actual process conditions.

4. Conclusions

Based on the F-test result, since the p-value was 0.000 (< 0.05) and $F_{count} > F_{table}$, it can be concluded that five coagulant feed water parameters simultaneously have a significant effect on the coagulant dose. Individual t-test result presented that turbidity, color, and conductivity had significant affect to coagulant doses ($p < 0.05$), in contrast pH and temperature didn't significant influence ($p > 0.5$). Empirical models were then developed using both all five parameters and a reduced set of the three statistically significant parameters. The ANN-BP models which have lowest MSE and highest R^2 were [5-6-1] ($R^2 = 0.96051$; MSE = 0.00179) and [3-4-1] ($R^2 = 0.97755$; MSE =0.00102). In conclusion, [3-4-1] is recommended because of the lowest MSE and highest R^2.

Acknowledgement

The authors would like to acknowledge the Institut Teknologi Sepuluh Nopember for the project funding under D4-2-ITS25 - Dana Departemen/Fakultas/Unit - Batch 2

References

[1] F. F. Pradypna, B. D. Marsono, and E. S. Soedjono, A Study of Drinking Water Supply and Demand in Surabaya in the Year 2039, IOP Conf. Ser.: Earth Environ. Sci. **506**, 012028 (2020).
[2] I. D. Joshua, A. C. Salihu, A. M. Mshelia, and N. N. Ubachukwu, ANALYSIS OF COMMUNITY-BASED PATTERN OF WATER DEMAND AND SUPPLY, Jessd **6**, (2023).
[3] M. Taufik, E. Khairina, R. Hidayat, R. Kalalinggi, and M. Iqbal, Study of Government's Strategy Indonesia on Clean Water Avaiability in Indonesia, Jurnal Kesehatan Lingkungan Indonesia (JKLI) **21**, 111 (2022).
[4] N. H. Pakharuddin, M. N. Fazly, S. H. Ahmad Sukari, K. Tho, and W. F. H. Zamri, Water treatment process using conventional and advanced methods: A comparative study of Malaysia and selected countries, IOP Conf. Ser.: Earth Environ. Sci. **880**, 012017 (2021).
[5] J. Bratby, Coagulation and Flocculation in Water and Wastewater Treatment, Water Intelligence Online **15**, 9781780407500 (2016).
[6] S. Narges, A. Ghorban, K. Hassan, and K. Mohammad, Prediction of the optimal dosage of coagulants in water treatment plants through developing models based on artificial neural network fuzzy inference system (ANFIS), J Environ Health Sci Engineer **19**, 1543 (2021).
[7] Z. Shi, C. W. K. Chow, R. Fabris, J. Liu, E. Sawade, and B. Jin, Determination of coagulant dosages for process control using online UV-vis spectra of raw water, Journal of Water Process Engineering **45**, 102526 (2022).
[8] E. L. L. Tochio, B. C. Do Nascimento, and S. R. Lautenschlager, Coagulant dosage prediction in the water treatment process, Water Supply **23**, 3515 (2023).
[9] M. M. Hammad, *Artificial Neural Network and Deep Learning: Fundamentals and Theory*, arXiv:2408.16002.
[10] Robert Odek and Gordon Opuodho, F-test and P-values: A synopsis, J.M.S **13**, 59 (2023).
[11] N. Dayarathne, M. J. Angove, S. Jeong, R. Aryal, S. R. Paudel, and B. Mainali, Effect of temperature on turbidity removal by coagulation: Sludge recirculation for rapid settling, Journal of Water Process Engineering **46**, 102559 (2022).
[12] Baghvand, Optimizing Coagulation Process for Low to High Turbidity Waters Using Aluminum and Iron Salts, American Journal of Environmental Sciences **6**, 442 (2010).
[13] R. Nedjai, A. Al-Mamun, and M. Z. Alam, Effects of initial turbidity and myco-coagulant dose on the effectiveness of the coagulation process in water treatment, Appl. Chem. Eng. **7**, 1546 (2024).
[14] A. Mortadi, E. G. Chahid, A. Elmelouky, M. Chahbi, N. El Ghyati, S. Zaim, O. Cherkaoui, and R. El Moznine, Complex electrical conductivity as a new technique to monitor the

coagulation-flocculation processes in the wastewater treatment of the textile Industry, Water Resources and Industry **24**, 100130 (2020).
[15] X. Ying, An Overview of Overfitting and its Solutions, J. Phys.: Conf. Ser. **1168**, 022022 (2019).