Comparison of Electrical Conductivity Prediction Models Using Gaussian Process

Zaenuri Putro Utomo[1], Indriana Hidayah[2], Muhammad Nur Rizal[3]

***Abstract*—People living in coastal areas use clean water sourced from groundwater to support the household, agricultural, and industrial needs. However, human activities and natural factors can lead to a common problem in coastal areas, namely seawater intrusion. Seawater intrusion can be detected using water quality data. Today, one of the challenges in water resources management is the prediction of water quality parameters such as total dissolved solids (TDS), electrical conductivity (EC), and water turbidity. Incomplete EC data and limitations of direct measurements can affect the analysis. Machine learning models are known to provide the most accurate predictions. This research used EC parameter data to investigate the performance of algorithms, namely artificial neural networks (ANN), Gaussian processes (GP), and multiple regression (MLR). The prediction used seven hydrochemical parameters (K, Ca, Mg, Na, SO4, Cl, HCO3) and three physical parameters of groundwater (TDS, pH, EC). Performance measurement used R-squared (R2) and root mean squared error (RMSE). The testing showed the MLR model had R2 of 0.985 and RMSE of 0.030, which were slightly better than other models. Hence, it can be concluded that the MLR model can be a solution to difficult problems of EC prediction and incomplete data in the water resources management.**

***Keywords*—Prediction, Electrical Conductivity, Water Quality, Groundwater.**

I. INTRODUCTION

Coastal and marine areas play a crucial role for the future of archipelagic countries, one of which is Indonesia. Stretching over 95,186 km, Indonesia becomes an archipelagic country with the longest tropical coastline in the world. Almost all of Indonesia's economic activities are in coastal areas as at least 75% of big cities and 80% of industries [1] are situated in these areas. Hence, coastal areas have an important role in the Indonesian economy, which, at same time, is also a threat to environmental resources. One of the issues in coastal areas is related to water quality, namely seawater intrusion which frequently occurs in shallow coastal aquifers [2]. Triggers can be caused by human activities, economic activities, and natural factors. Therefore, seawater intrusion becomes an essential study for coastal cities.

An in-depth study of the sea water intrusion phenomenon is needed for policymaking in water management activities in coastal areas. Seawater intrusion can be detected by measuring salinity. Traditionally, salinity measurement uses monitoring wells along coastal areas. The measurement results can be modeled and used for monitoring groundwater management. Complete water quality data in terms of the number of parameters as well as an even distribution of locations can be a support in monitoring and planning activities.

Data availability and validity is an obstacle in groundwater management activities. Groundwater in Semarang City has been exploited since 1841 and continues to increase sharply [3] with the construction of deep wells. After 2000, groundwater utilization through the construction of deep wells reached 1,194 locations. Increased utilization of groundwater can be a factor causing seawater intrusion. Water is a major concern and a basic need in life. Water quality provides an overview of safety and suitability for its intended purpose. Measurement of water quality consists of several factors, including physical, biological, and chemical water.

Hydrochemical data or groundwater chemical water has been studied by researchers from various fields. Soil chemical conditions are often associated with physical parameters. One of the most informative, important, and easy technique to measure water quality parameters for salinity is electrical conductivity (EC). EC provides information quickly and precisely about the condition of the electrolyte [4] in the water. Reference [5] is one example of a study examining pollution in a mining area. The study has shown that the increase in the EC concentration indicates the level of pollution in surface water and ground water.

Groundwater sampling collected from observation wells must be immediately measured to obtain the EC value. Direct measurement is mandatory because changes in water temperature affect the actual value of EC. Sampling and direct measurements on-site require time, cost, and effort [5]–[7]. The difficulty of reaching the location is added to the list of considerations for choosing direct measurement as a solution to complete the data. Reference [2], [3] illustrated that at a certain time, relatively little data was collected in comparison to the vast area covered. In other cases, there were many missing water quality data. Incomplete water quality data is a serious issue that must be addressed. Incomplete water quality data sets or gaps at certain locations or time periods require efforts to complete them.

Today, one of the most difficult issues in water resource management is predicting water quality parameters such as TDS, EC, and turbidity [8]. Reference [9] stated that EC was one of the core parameters on water quality assessment. The other previous studies used machine learning (ML) for the prediction [4], [5], [7] or seawater intrusion modeling [2], [10], completeness of water quality data [3], and direct measurement constraints [4], [6]. There are three problems related to groundwater quality, including prediction or predictive models, incomplete data, and difficulties in direct measurement. Based on these three problems and literature review on the EC

[1,2,3] *Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jln. Grafika No. 2, Kampus UGM, Yogyakarta, 55281, INDONESIA (Tel. +62-274-552305, email: [1]zaenuri.p.u@ugm.ac.id, [2]indriana.h@ugm.ac.id, [3]mnrizal@ugm.ac.id)*

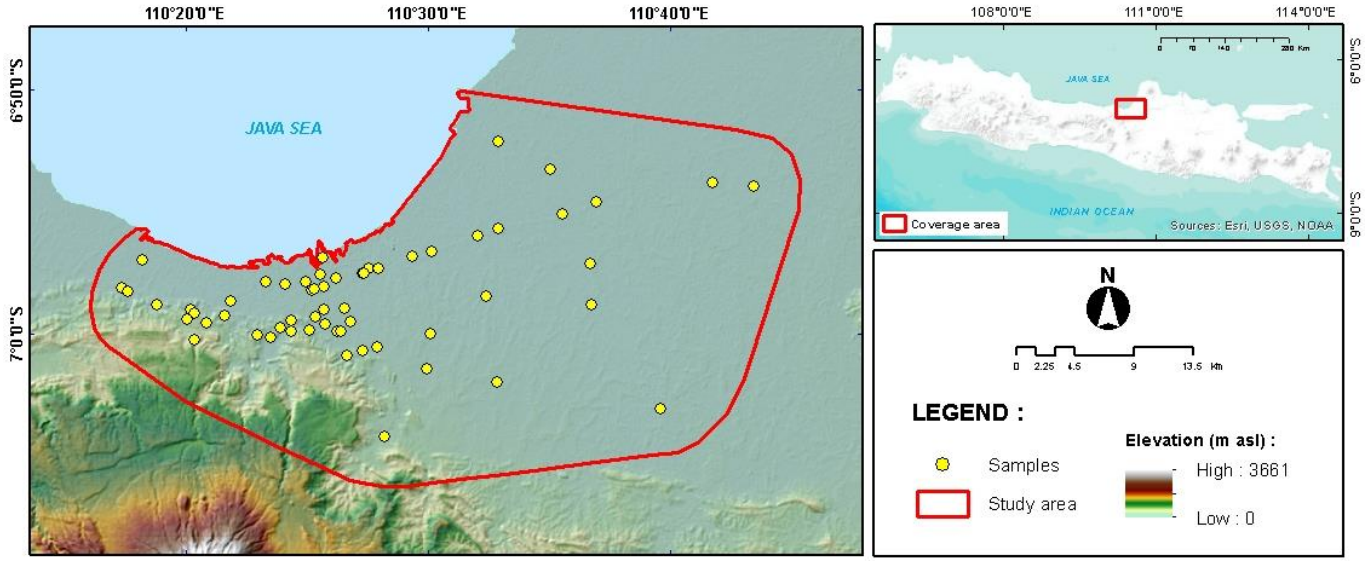

Fig. 1 Map of study area and sample of wells in Semarang-Demak.

prediction, this study tested several machine learning algorithms to develop an accurate, reliable, and efficient EC prediction model. This study measured the performance prediction of the artificial neural network (ANN), gaussian process (GP), and multiple linear regression (MLR) algorithms on groundwater datasets. The accuracy and time required to build a model are used to evaluate performance. In [7], [11], [12], the performance was measured using two tools, namely root mean squared error (RMSE) and R-squared (R^2). This paper assessed EC as a key parameter so that a model was needed as an alternative in solving the problem. The best performing algorithm was proposed as an EC prediction model. In addition to the prediction problem, the model was a solution to the lack of water quality data (especially EC) and the constraint of direct measurement.

II. Related Work

EC prediction is a concern for many studies [2], [4], [5], [7], [11], [12] by providing accurate solutions. Reference [7] conducted a study and evaluation of water quality using parameters of water discharge, pH, temperature, Na, Ca, and Mg. The results compared ANN and MLR. The experimental data were compared with the values generated by the ANN. The ANN model generated accurate results in one hidden layer with seven neurons. The proportional contribution percentage of each input variable was determined using Garson's approach of network connection weights. Variables of temperature and the amount of concentration of Ca plus Mg have the most influence on the prediction of EC. The results indicate that the ANN model outperforms the MLR model. Although the two proposed models are capable of approaching the EC parameters very well, the time required to develop the model is not described in this study.

The same algorithm has been carried out but with different parameters [5]. EC prediction used eleven parameters as input in the analysis, including K, Na, Ca, Mg, Cl, SO_4, HCO_3+CO_3, Fe, Mn, Al, and NO_3. Similar to the algorithm in previous studies [5], [12] in obtaining the most suitable ANN and MLR architecture, this study employed Rapidminer 9.9 by testing on several hidden layers, activation functions, number of neurons, optimizer, and the resulting prediction model. The results demonstrated that the machine learning model was applicable and capable of providing high accuracy and reliability for EC. Since the number of hidden layers and neurons differed between the table and the results, a back test between the differences was required.

Reference [12] estimated the EC of groundwater samples in northern Gallikos River, Greece. Between 2004 and 2005, the capacity of ANN and MLR to handle 233 samples from 89 sampling points was evaluated. To identify the input water quality metrics, a Pearson correlation matrix and factor analysis were used. Ca, Mg, Na, and Cl were used as input parameters. A model with a single hidden layer of three neurons produced the best ANN outcomes. The results indicated that while both ANN and MLR models performed equally well with larger samples (with the same input), MLR performed better with smaller data sets. The RMSE value was displayed in the result table as a fairly large validation value. The ANN and MLR algorithms were retested by considering the discrepancies found in previous studies [6], [11].

The Gaussian process (GP) is one of the machine learning algorithms. The GP is a nonparametric machine learning technique used to construct probabilistic models. It was employed to estimate NO_3^- and Sr contents in a groundwater quality database [13]. The M5P, random forest (RF), and random tree (RT) algorithms were compared to GP in terms of performance. Water quality parameters that are used as input to the model include temperature, pH, EC, Ca, Na, K, Mg, Cl, SO4, HCO3, and F. A tenfold cross-validation technique was employed to model the two subgroups. The data set was divided into two sections in which 173 samples were used to create the model and 73 samples were used to validate the model. The resulting model showed that the GP algorithm was superior to other models in predicting NO_3^- and Sr.

TABLE I
STATISTICAL DESCRIPTION OF SEMARANG DATASET

Parameter	Training (0.67; 39 sample)				Testing (0.33; 19 sample)			
	Min	Max	Avg	Std	Min	Max	Avg	Std
TDS	152	15,947	1,170	2,515	231	1,960	775	494
pH	6.4	11.6	7.5	0.8	6.5	8.7	7.4	0.6
EC	226	23,900	1,754	3,767	349	3,010	1,166	739
K	2	44	11	8	3	49	10	10
Ca	3	730	70	131	4	197	50	48
Mg	1	1,812	70	288	4	72	20	17
Na	26	7,200	354	1,135	22	580	186	166
SO_4	6	163	40	34	12	166	52	38
Cl	13	15,753	630	2,511	11	1,242	223	341
HCO_3	79	2,263	353	340	130	520	304	118

After conducting a literature analysis on groundwater quality prediction based on EC factors, this study used [5], [7], [12] as references. The ANN and MLR algorithms were used in all three investigations. Using the ANN method and seven neurons in one hidden layer, [7] proposed an EC prediction model. The first model employed the ANN algorithm with one hidden layer and three neurons, as described in [12]. The second model used the ANN method with one hidden layer and three neurons, as described in reference [13]. The MLR method was used in the second model. As in [5], [7], [11], [12], the model was given water quality parameters. Among the indicators of water quality were TDS, pH, EC, K, Ca, Na, Mg, Cl, SO_4, HCO_3. The prediction values produced by both models were nearly identical. When dealing with small amounts of data, ANN outperformed MLR. Meanwhile, the MLR model outperformed the ANN model when dealing with massive amounts of data. The anticipated values of ANN with 50.81 and MLR with 49.74 are shown in the RMSE value calculated in [7]. ANN precision values of 175.9 and MLR of 168 were obtained from [12]. The RMSE is a metric for evaluating a predictive model's performance; the lower the value, the better the model's performance [12].

To overcome the large RMSE value in the preprocessing stage, this study presents a normalizing strategy. Reference [5] has provided two prediction models, one with nine neurons and the other with three hidden layers using a 7-3-16 neuron configuration. Between the description (7-3-16) and the table, the ANN model with three hidden layers revealed the number of distinct neurons (7-13-16). The architecture in question was determined by testing the difference. Reference [5], [7], and [12] has built a foundation for this research to reexamine the ANN and MLR algorithm models. GP is another proposed algorithm. Reference [6] used the GP algorithm to develop a nitrate and strontium prediction model that outperformed the M5P, RF, and RT.

Previous study has shown that gaps in problems could predict EC. As a result, an inquiry is required to develop a model that can forecast the EC value accurately, reliably, and quickly. This research compared the predicted results of GP, ANN, and MLR on groundwater data sets. In addition to the time required, the prediction model was tested using two measuring tools as in previous studies, including R^2 and RMSE. The purpose of this research is to compare the predictions of GP, ANN, and MLR on groundwater data sets. Along with the time necessary, the prediction model was validated using the same two metrics used in earlier studies, namely R^2 and RMSE.

III. MATERIAL AND METHODS

A. Study Area

Semarang-Demak is part of a coastal area on the island of Java, Indonesia. The research area, extending from Semarang to Demak, is located on Java's northern shore covering an area of 1,070 km^2 and is home to about three million people. It is located between 419,500 to 480,250 m in east longitude and 9,212,850 to 9,258,190 m south latitude by using Universal Transverse Mercator (UTM) coordinate system as in Fig. 1. Groundwater exploitation or drilling has been begun since 1841 when deep wells were constructed [3]. Since then, the number of deep wells has increased increasingly. This condition indicates that the need for groundwater is very essential, both for household and industrial purposes. In addition to quantity, good groundwater quality must be an immense attention. This research is part of an effort to complete the water quality data set.

B. Data Used

The data set was secondary data from sample measurements in several regions of the country. Semarang dataset [3] was the primary dataset, providing the hydrochemical parameters of groundwater samples collected near Semarang, Central Java, Indonesia. In 1992, 1993, 2003, 2006, and 2007, a total of 58 samples were gathered based on drill point data from multiple Ministry of Energy and Natural Resources reports and independent consultants. The set consisted of parameters: X coordinate, Y coordinate, well depth, water level, total dissolved solids or TDS, pH, EC, K, Ca, Na, Mg, Cl, SO_4, HCO_3, year, ion balance, aquifer, and hydrogeochemical facies. The chemical composition was tested at the Diponegoro University Water Quality Institute using spectrophotometry. This study did not use all of the parameters in the dataset. The parameters used included seven groundwater hydrochemical parameters (K, Ca, Mg, Na, SO_4, Cl, and HCO_3), three physical parameters (TDS, EC, and pH) to make EC prediction models.

Table I provides statistical description of the Semarang dataset. Some water quality parameters have different units. The pH parameter has no units, while the EC is µS/cm. The units are meq/L for TDS, K, Ca, Mg, Na, SO4, Cl, and HCO_3 parameters.

The size of the data set can affect the test results and the performance of an algorithm in modeling. Therefore, the resulting model was tested using several sources of public datasets of different sizes and sources. Other datasets used [5], [6], [14] had 60 rows, while [15] had 799 rows. The dataset had three physical parameters namely EC, TDS and pH. The other seven parameters were groundwater hydrochemistry (Na, K, Ca, Mg, HCO_3, Cl, and SO_4).

The whole process was divided into three parts. Fig. 2 illustrates the process of developing an EC prediction model. The first stage was data set and preprocessing. At this stage, each dataset selected seven hydrochemical parameters and two physical groundwater parameters. Overall, each water quality parameter has a different unit. Therefore, the initial process was to normalize the dataset. The dataset was divided into two sections in which 67% used for training and 33% for testing. The main part of the EC model development process was to compare the performance of the ANN, GP, and MLR algorithms. In the final part, the model was evaluated to get the best performance in the EC prediction. The evaluation used three measuring tools, including RMSE, R^2, and model development time.

C. Identification of Optimum Value

The next step was to determine the optimal value for each model. After that, it is applied to the operators of each model by repeating trial and error [5], [12]. Rapidminer Studio 9.9 was used for all calculations. Each algorithm was checked to find its optimal parameters. As for ANN, the settings of [5] was used so that it could be compared. The first step was to run the operator on the first iteration using the default values. The results obtained were then evaluated and considered for a better value in the next process. The mechanism was repeated until the optimum value for each operator model was obtained. After each operator got the optimal value in the dataset, the resulting model was evaluated using a validation model.

After proper configuration to get the optimal value of each algorithm, seven predictive models were used to estimate one of the main water quality parameters, namely EC. The seven models were built from the ANN, GP and MLR algorithms. An overview of the method behavior will be described in this section. The EC prediction model will be built using the following algorithm:

1) Gaussian Process (GP): GP is a stochastic process that is Bayesian-analyzed using a simple matrix [16]. GP is an algorithm type that is based on the normal distribution. It is a general-purpose machine learning algorithm that may be used to address problems involving classification and regression. The GP kernel functions can be utilized to predict unseen values from the training data in regression problems [6]. GP prediction with high accuracy depends on the suitability of the selected kernel [17]. Therefore, the search for optimal EC values will involve kernel type and length scale.

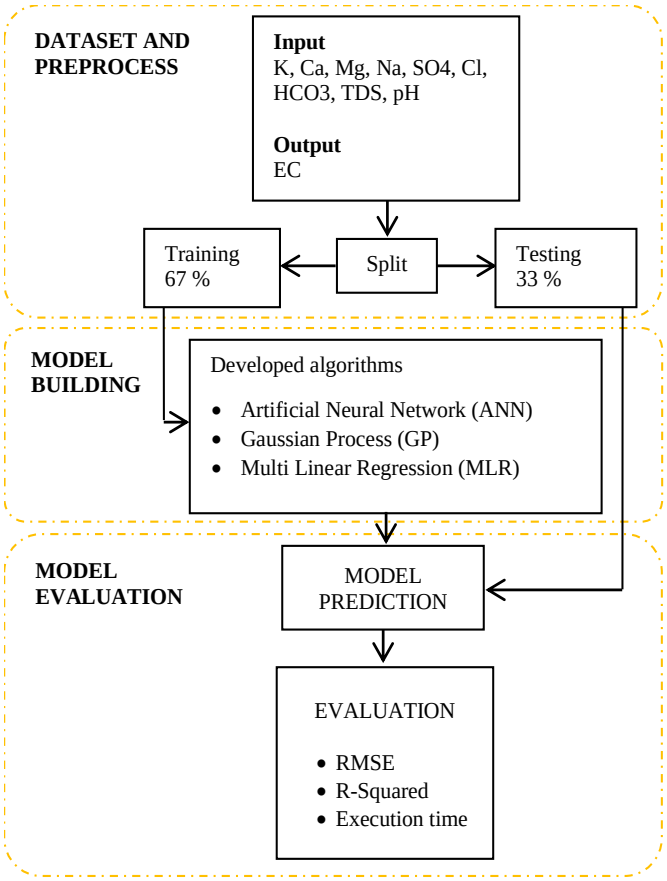

Fig. 2 EC prediction process overview.

2) Artificial Neural Network (ANN): ANN architecture consists of an input layer, a hidden layer, and an output layer. The ANN algorithm is a technique like the problem-solving process of the human brain. The ANN model employed one and three hidden layers based on reference [5]. Nine neurons were located in one hidden layer, whereas the two models with three hidden layers had 7-3-16 and 7-13-16 neurons. The data was divided into 67% training and 33% testing. The hidden layer settings and test size in [5] was used in this research. In addition to the three models, the previous ANN architecture, namely one hidden layer with seven neurons [7] and one hidden layer with three neurons [12] were included in the test.

3) Multiple Linear Regression (MLR): MLR is a generally used and precise technique [16] that builds meaningful equations between the dependent variable and a set of independent variables that serve as predictors. Reference [16], [17] have utilized this method successfully in hydrogeology and hydrochemistry to forecast water quality and construct statistical models. Equation (1) illustrates the calculations on the MLR algorithm.

$$y = a_0 + \sum_{i=1}^{n} a_i X_i. \quad (1)$$

Refer to (1), y is the dependent variable or output, a_0 is a constant, a_i is the regression coefficient for the independent variable i, and X. In this research, MLR provided equations in EC prediction. It is the advantage of MLR [12] over ANN. The

TABLE II
PERFORMANCE EVALUATION OF EC PREDICTION MODELS

Dataset	Model	RMSE	R2	Execution Time (m:s)
Semarang	ANN_3	0.049	0.955	00:00
	ANN_7	0.045	0.963	00:00
	ANN_9	0.067	0.967	00:00
	ANN_7-3-16	0.047	0.966	00:00
	ANN_7-13-16	0.041	0.971	00:00
	GP	0.298	0.718	00:00
	MLR	0.030	0.985	00:00
Keskin & Özler	ANN_3	0.036	0.999	00:00
	ANN_7	0.107	0.988	00:00
	ANN_9	0.129	0.984	00:00
	ANN_7-3-16	0.397	0.858	00:00
	ANN_7-13-16	0.722	0.656	00:00
	GP	0.976	0.484	00:00
	MLR	0.009	1.000	00:00
Beetaloo	ANN_3	0.277	0.851	00:08
	ANN_7	0.347	0.790	00:17
	ANN_9	0.480	0.639	00:22
	ANN_7-3-16	0.245	0.900	00:45
	ANN_7-13-16	2.864	0.132	01:09
	GP	0.698	0.965	00:00
	MLR	0.116	0.973	00:00

MLR capability was compared with other algorithms, namely GP and ANN.

D. Performance Evaluation of the Models

Two measures were used to evaluate the model's performance, namely RMSE and R^2. RMSE is a tool for evaluating linear regression models by determining the level of accuracy of a model's prediction results. RMSE is calculated by squaring the error (prediction – measurement) divided by the amount of data (= average), then taking the root. RMSE has no units. Equation (2) is the formula for RMSE. R^2 or coefficient of determination has a function to measure the success of the regression model in predicting the value of the dependent or dependent variable. Referring to (3), R^2 assesses the effect of independent or independent variables together on the dependent variable. The formulas of the two measuring tools are as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(K_p - K_m\right)^2} \quad (2)$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(K_p-\bar{K}_p)(K_m-\bar{K}_m)}{\sqrt{\sum_{i=1}^{n}(K_p-\bar{K}_p)(K_m-\bar{K}_m)}}\right)^2 \quad (3)$$

with

n : number of samples
K_p : predictive value
K_m : measurement value

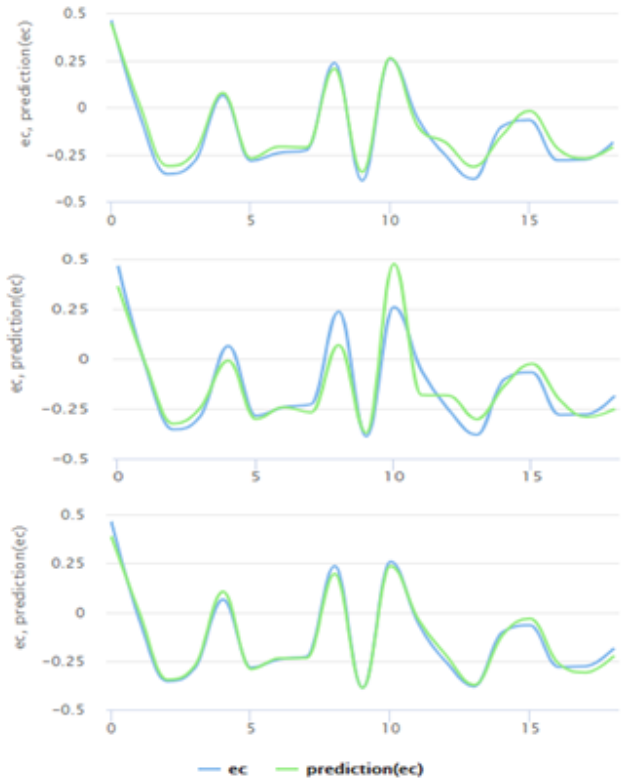

Fig. 3 Comparison graph of EC concentration observation and prediction results algorithm (top to bottom) ANN, GP and MLR.

$\bar{K}_p$: mean value of prediction
$\bar{K}_m$: mean value of measurement.

The model with high R^2 and low RMSE showed good performance [12].

IV. RESULTS AND DISCUSSIONS

The optimal value of each algorithm in handling water quality data was conducted by trial-and-error method. The performance of the algorithm for obtaining EC values from water quality datasets was measured using RMSE and R^2. Three public data sets were used to compare the reliability of the prediction models. In addition to performance, the time required to create a model was also measured. EC prediction experiments using three algorithms are presented in Table II. The findings of optimal value of three algorithms are presented in Fig. 3. The comparison between the measured (blue line) and predicted (green line) values suggests that the MLR algorithm provides the best predictive results. The prediction line shows a small deviation compared to the ANN and GP algorithms.

Specifically, for the ANN algorithm, there were five experiments with different hidden layers. The architecture of the number of layers and ANN neurons was the same as in [5], [7], [12]. Table II shows the ANN algorithm with one hidden layer such as ANN_3, ANN_7, and ANN_9. It is considered necessary to be reexamined because there are differences in the configuration of the three hidden layers between the table and the description. The table listed neurons 7-13-16, but the discussion of the data referred to neurons 7-3-16.

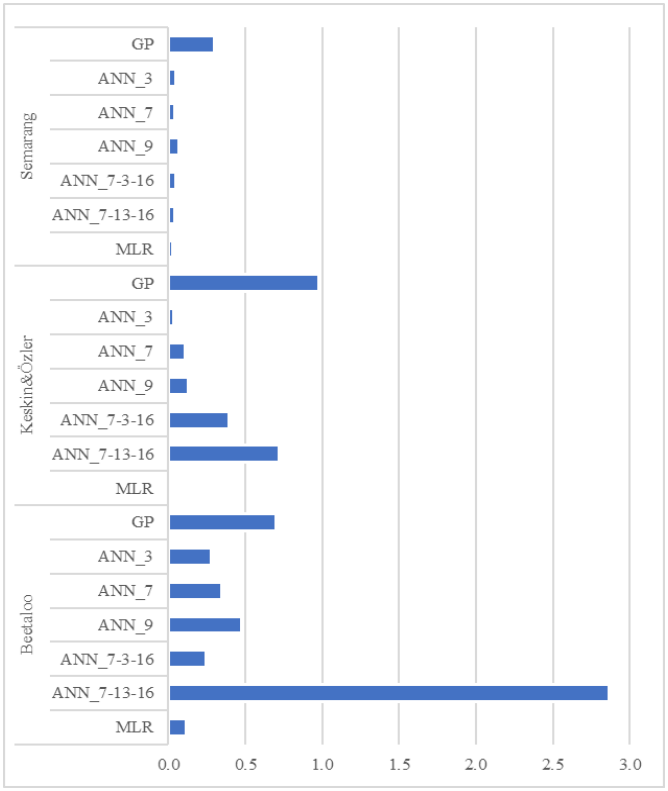

Fig. 4 Bar chart of EC Prediction Model Performance based on RMSE.

Each dataset of modeling was conducted seven times, five-time ANN and one time for each GP and MLR algorithm. In the first dataset [3], the MLR model generated the lowest RMSE value of 0.030, followed by ANN models and GP. In the same order, the MLR model outperformed the other models by measuring R^2, which was 0.985. All models using the first dataset took less than 1 second. The prediction model with the first dataset indicates that the MLR model has better performance than the other six models. The EC prediction model using the dataset [6] (size 60 rows) had slightly different results than before. The performance of MLR and ANNs with one hidden layer was better than other models. The order of the best model performance using the second dataset was the same between RMSE and R^2. The MLR model obtained an RMSE value of 0.009 and a value of 0.036 for ANN_3. Slightly different in terms of execution time with the first dataset, the whole model took less than 1 second to build the model. The third dataset [15] had 799 rows, larger than the two previous datasets. The EC prediction model tested using large data showed different results except for the MLR performance which remained superior to other models. Based on the RMSE value, the best performance after MLR was the ANN model with one hidden layer, ANN_7-3-16, GP, and ANN_7-13-16. In contrast to R^2, the order of the best performance after MLR was GP, ANN_7-3-16, ANN_9, and ANN_7-13-16.

The following description based on Table II discusses the performance of the model in relation to the three datasets. The performance measure used RMSE, R^2, and execution time. In general, the ANN model's performance was better than the GP, even though its execution time was slower. The best performing ANN was a model with one hidden layer with seven neurons then three neurons. A configuration with three hidden layers, ANN_7-3-16 performed better than ANN_7-13-16. These results can support [6], which has revealed that the ANN with three hidden layers with the best performance on the EC prediction model is ANN_7-3-16.

In contrast to the results in [6], argued that the performance of ANN was slightly better than MLR, this experiment suggests that the best performance is the MLR model. Reference [11] has stated the same thing, that MLR is slightly better than the ANN model. Even though GP's performance was not the best, it showed a good performance in terms of record time, which was under 1 second. The EC measurements compared with the predicted results in bar chart can be seen in Fig. 4. MLR gives better predictive results than other algorithms. In line with the results of Table II with a large data size [15], a number of 799 rows show a better predictive value than other algorithms. Consideration of the value of RMSE, R^2, and execution time showed that the MLR and ANN (with a hidden layer) algorithms yielded good results for the three datasets. When compared to the prediction capabilities of the three datasets, MLR was superior and required an execution time of less than 1 second. The MLR model performed reliably, accurately, and quickly regardless of the size of the data or the number of rows (up to 799 rows). The model created equations, which was another advantage of the MLR approach over ANN. The following equations are provided by the MLR model used in this research.

$$ECp1 = 1.033 \times \mathrm{TDS} - 0.011 \times \mathrm{Ca} + 0.099 \times \mathrm{Mg} - 0.141 \times \mathrm{Cl} + 0.021 \times \mathrm{HCO3} \quad (4)$$

$$ECp2 = 1.001 \times \mathrm{TDS} - 0.005 \times \mathrm{K} - 0.010 \times \mathrm{Ca} - 0.007 \times \mathrm{Mg} - 0.004 \times \mathrm{Na} + 0.004 \times \mathrm{SO4} + 0.007 \times \mathrm{Cl} + 0.017 \times \mathrm{HCO3} \quad (5)$$

$$ECp3 = 0.663 \times \mathrm{TDS} + 0.114 \times \mathrm{Ca} + 0.256 \times \mathrm{Na} - 0.146 \times \mathrm{SO4} \quad (6)$$

From the three datasets, three equations were created. Equation predicted the EC value of the first dataset as ECp1 (4). TDS, Ca, Mg, Cl, and HCO3 were water quality characteristics influencing the predicted EC concentration. The projected value of ECp2 was given by (5) in the second dataset. The water quality factors TDS, K, Ca, Mg, Na, SO4, Cl, and HCO3 influenced the projected value of EC. The anticipated value of ECp3 was obtained from (6), from the third dataset. TDS, Ca, Na, and SO4 are all factors that influence the expected value of EC. The TDS and Ca parameters had an impact on the projected values in the MLR model's three equations. The TDS and Ca parameters both contribute to the projected value of EC, as seen in the three equations. The EC value was strongly influenced by the TDS parameter.

V. Conclusion

EC is one of the key parameters in water quality evaluation. Obstacles in direct measurement, lack of data, and predictions

are problems in water resource management. A model of EC forecasts based on water quality data is important for regional analysis and planning. The research findings showed that the MLR algorithms had good ability to predict EC values. According to the MLR results, R^2 was greater than 0.970, RMSE was between 0.009 and 0.116, and the time execution was under 1 second. MLR was slightly superior to ANN. The model had a good ANN performance score; however, the larger the data, the longer it took to build the model. The ANN models with one hidden layer could be an alternative when time was not a mandatory requirement. In addition, the MLR model for EC prediction can be a solution to the problem of groundwater quality datasets. Strong predictive ability and a fairly short time are needed to build a water resource management system.

ACKNOWLEDGMENT

This research was supported by master's degree scholarship program The Ministry of Communication and Information Technology.

REFERENCES

[1] R.R. Sigit (2013) "Tahun 2020, Pemerintah Targetkan 20 juta Hektar Kawasan Konservasi Perairan dan Laut," [Online], http://www.mongabay.co.id/2013/07/10/tahun-2020-pemerintah-targetkan-20-juta-hektar-kawasan-konservasi-perairan-dan-laut/, access date: Apr. 7, 2021.

[2] C. Sivapragasam, V. Jegatheesan, V.M. Arun, and S. Vanitha, "Spatial Modeling of Electrical Conductivity with Neural Network," *International Journal of Engineering Science and Technology*, Vol. 2, No. 7, pp. 3128–3136, 2010.

[3] T.T. Putranto and T.R. Rüde, "Hydrogeological Model of an Urban City In a Coastal Area, Case Study: Semarang, Indonesia," *Indonesian Journal on Geoscience*, Vol. 3, No. 1, pp. 17–27, Apr. 2016.

[4] N.A. Memon, M.A. Unar, A.K. Ansari, G.B. Khaskheli, and B.A. Memon, "Predictive Potentiality of Artificial Neural Networks for predicting the Electrical Conductivity (EC) of Drinking Water of Hyderabad City," *ICCOMP'08: Proceedings of the 12th WSEAS international conference on Computers*, 2008, pp. 487–490.

[5] T.E. Keskin, E. Özler, E. Şander, M. Düğenci, and M.Y. Ahmed, "Prediction of Electrical Conductivity Using ANN and MLR: a Case Study from Turkey," *Acta Geophysica*, Vol. 68, No. 3, pp. 811–820, Jun. 2020.

[6] D.T. Bui, K. Khosravi, M. Karimi, G. Busico, et al., "Enhancing Nitrate and Strontium Concentration Prediction in Groundwater by Using New Data Mining Algorithm," *Science of The Total Environment*, Vol. 715, pp. 1–13, May 2020.

[7] M.A. Ghorbani, M.T. Aalami, and L. Naghipour, "Use of Artificial Neural Networks for Electrical Conductivity Modeling in Asi River," *Applied Water Science*, Vol. 7, No. 4, pp. 1761–1772, Jul. 2017.

[8] I. Ahmadianfar, M. Jamei, and X. Chu, "A Novel Hybrid Wavelet-Locally Weighted Linear Regression (W-LWLR) Model for Electrical Conductivity (EC) Prediction in Surface Water," *Journal of Contaminant Hydrology*, Vol. 232, pp. 1–17, Jun. 2020.

[9] Friends of Groundwater, "Assessing Groundwater Quality: A Global Perspective: Importance, Methods and Potential Data Sources," World Water Quality Alliance, Nairobi, Kenya, Report, 2021.

[10] R.S.B. Waspodo, S. Kusumarini, and V.A.K. Dewi, "Prediksi Intrusi Air Laut Berdasarkan Nilai Daya Hantar Listrik dan Total Dissolved Solid di Kabupaten Tangerang Prediction," *Jurnal Teknik Pertanian Lampung*, Vol. 8, No. 4, pp. 234–303, Dec. 2019.

[11] B. Tutmez, Z. Hatipoglu, and U. Kaymak, "Modelling Electrical Conductivity of Groundwater Using an Adaptive Neuro-Fuzzy Inference System," *Computers & Geosciences*, Vol. 32, No. 4, pp. 421–433, May 2006.

[12] C. Mattas, L. Dimitraki, P. Georgiou, and P. Venetsanou, "Use of Factor Analysis (FA), Artificial Neural Networks (ANNs), and Multiple Linear Regression (MLR) for Electrical Conductivity Prediction in Aquifers in the Gallikos River Basin, Northern Greece," *Hydrology*, Vol. 8, No. 3, pp. 1–,14 Sep. 2021.

[13] G.S. Bhunia, A. Keshavarzi, P.K. Shit, E.-S.E. Omran, and A. Bagherzadeh, "Evaluation of Groundwater Quality and Its Suitability for Drinking and Irrigation Using GIS and Geostatistics Techniques in Semiarid Region of Neyshabur, Iran," *Applied Water Science*, Vol. 8, No. 6, pp. 1–16, Oct. 2018.

[14] T.E. Keskin and E. Özler, "Heavy Metal Contamination in Groundwater and Surface Water due to Active Pb-Zn-Cu Mine Tails and Water-Rock Interactions: A Case Study from the Küre Mine Area (Turkey)," *Turkish Journal of Earth Sciences*, Vol. 29, No. 6, pp. 878–895, Nov. 2020.

[15] *Groundwater and Surface Water Sample Locations Included in Hydrochemical Cluster Analysis*, Geological and Bioregional Assessment Program, May 2020, [Online], https://data.gov.au/data/dataset/27b0fbba-5a68-4055-8600-c181dac15ffb, access date: Dec. 18, 2020.

[16] A.J. Smola and P. Bartlett, "Sparse Greedy Gaussian Process Regression," *NIPS'00: Proceedings of the 13th International Conference on Neural Information Processing Systems*, 2000, pp. 598–604.

[17] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*. London, UK: The MIT Press, 2012.