International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1408 - 1420
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0194

Prediction of Optimum Dosage of Coagulant in Water Treatment Plant: A Comparative Study between Artificial Neural Network and Random Forest

Nitin T. Sawalkar[1], *Sagar W. Jadhav*[2], *Alpita A. Pawar*[3]
[1,2,3] *Department of Civil Engineering, VIIT, Pune-MH, 411048, India*
Emails: *nitin.sawalkar@viit.ac.in*[1], *sagar.22120002@viit.ac.in*[2], *alpita.22120096@viit.ac.in*[3]

Abstract

Raw water, sourced directly from natural water bodies, is unsuitable for direct consumption due to the presence of various impurities. Therefore, it undergoes treatment at a Water Treatment Plant (WTP) before being supplied to the public. Preliminary treatment involves the removal of floating matter, through screening, while heavier particles settle out by gravity, fine particles remain in suspension, causing turbidity. Effective removal of these suspended particles requires coagulation to form flocs and facilitate the settling. Determining the optimal coagulant dosage is crucial, as both underdoing and overdosing of coagulant can lead to ineffective treatment and increased costs. Conventionally optimum dosage of coagulant is determined by performing jar test. This study focuses on predicting the optimum coagulant dosage using two soft computing techniques: Artificial Neural Network (ANN) and Random Forest (RF). The Input parameters for model development include turbidity, pH, temperature, and alkalinity of raw water from the Parvati Water Treatment Plant, Pune. In this study Four models were developed, namely Model A (Turbidity), Model B (pH, Alkalinity, Temperature, Turbidity), Model C (pH, Alkalinity, Temperature), and Model D (Alkalinity and Turbidity). These models were trained using ANN and RF. Predictions of optimum coagulant doses were made for the testing dataset, and model accuracy was evaluated using Scatter plots, Root Mean Squared Error (RMSE) and Coefficient of Correlation (R). Results indicate that RMSE values of ANN Models are comparatively lower than RF. Comparing among Models A, B, C, and D, Model B and Model D exhibit better performance, with lower RMSE values.

Keywords: *Artificial neural network; Random forest; Soft computing; Water treatment;*

1. Introduction

Water sourced through natural water bodies is unsuitable for direct consumption due to various impurities present. Water treatment is an essential process that helps ensure the potability of water by removing the impurities present in water. There are mainly three types of solids present in water suspended and dissolved. The potability of water is mainly dependent on Physical, Chemical and Biological properties of water. The potability of water is governed by Bureau of Indian Standards 10500 (2012). The quality of water has to be in a way that it satisfies the BIS limits given for Physical, Chemical, Biological properties of water. Conventionally, the coagulant dose needed for coagulation is determined empirically through laboratory jar test, where a single test may take at least 1 hour to be performed; during period of fast variation of water characteristics (e.g., during floods) it is impossible to have a real-time response. Moreover, this conventional technique, due to manual intervention, can lead to excessive or insufficient non-adequate coagulant doses [1]. To address these challenges, researcher have turned to soft computing techniques for more accurate and dynamic coagulant dose prediction. These techniques enhance the efficiency and effectiveness of water treatment process, ultimately improving the quality and safety of treated water. There have been various studies using such soft computing techniques for prediction of optimum dosage of coagulant. Artificial Neural Network (ANN) is one of the soft computing methods, many researchers have analysed the quality of water and have built various models to predict the different parameters related to water quality. The study aimed to compare Random Forest and Artificial Neural Network to predict the optimal dose of coagulant required for water treatment. The objective of this study is to compare developed models to determine the most suitable technique for prediction of coagulant and the most influential input parameters for model development

2. Literature Survey

Optimum dose of coagulant is very important parameter taken into consideration in water treatment process. Many of the researchers have worked to find out the optimum dosage of coagulant by using soft computing techniques. The literature survey of the same is given below.

The study by Baouab et al. (2018) aimed for predicting the optimal dose of coagulant in various potable water treatment processes. study considered several key parameters such as raw water turbidity, pH, temperature, alkalinity, total dissolved solids and hardness. The ANN models were trained using the Levenberg-Marquardt algorithm, the research findings suggest that ANN's can accurately predict the optimal coagulant dosage, offering a promising approach to optimizing water treatment processes [1]. Prediction of turbidity and aluminium in drinking water treatment plants using Hybrid Network (GA-ANN) and GEP was done by Alsaeed et al. (2021). The authors developed a hybrid network model combining genetic algorithms (GA) and artificial neural network (ANN) to predict turbidity and aluminium concentration in drinking water treatment plants. The study utilized input variables such as raw water quality parameters, coagulant type and dosage, and treatment process variables to develop the hybrid GA-ANN model. The results showed that the hybrid model was effective in predicting turbidity and aluminium concentration, and the study suggested that the model could be used as a tool to optimize coagulant dosage and improve water treatment performance [2]. Kote et al. (2019) have done modelling of chlorine and coagulant dose in a water treatment plant by artificial neural network. The study focused on using input parameters such as turbidity, total dissolved solids, pH and electrical conductivity to train and validate ANN models. Experimental data were used in this process. The study's results indicate that using ANN model can accurately predict the optimal coagulant dosage for water treatment plants [3].

The study by A. B. Sengul and Z. Gormez (2013) focused on the prediction of optimal coagulant dosage in drinking water treatment using Artificial Neural Network (ANN). In this study operational data from a drinking water treatment plant in Istanbul was used. The researchers created multiple ANN models. These models were trained using data on raw water quality parameters and alum dosage, is able to accurately predict the pH, turbidity, conductivity, color, UV254 and alum dosage of treated water are used. The effectiveness of model in predicting these parameters, the conductivity and pH, which are key indicators of water quality is assessed in the study [4]. Modelling and optimization of coagulant dosage in water treatment plants using hybridized random forest (RF) model with genetic algorithm (GA) optimization by Mohammed Achite et al. (2023) focuses on the Sidi Yacoub Water Treatment Plant in Algeria. Two models, RF and hybrid GA-RF model, have been created and compared. Different input scenarios are being examined to determine the best combination of input parameters for both models. The results indicate that GA-RF model, include raw water production, turbidity, conductivity and suspended material as input parameters outperformed the standalone RF model. This study helps improve coagulant dosage optimization in WTPs, leading to better operational efficiency [5]. Study by Dongsheng Wang et al. (2023) focuses on prediction for coagulant dosage and effluent turbidity of a coagulation process in a drinking water treatment plant based on the Elman Neural Network (ENN) and Random Forest (RF) models. This study focuses on addressing the challenges posed by the unpredictability in raw water quality and the time lag in the coagulation process. To overcome these challenges, the study proposes the use of two predictive models: The ENN and The RF model [6]. Study by Salim Heddam et al. (2021) used extremely randomized tree for predicting coagulant dosage in drinking water treatment plant. This study proposes and compare two soft computing techniques-extremely randomized tree (ERT) and random forest (RF) models for predicting coagulant dosage. These models are created using important water quality factors (such as turbidity, pH, dissolved oxygen, electrical conductivity, and water temperature) as inputs. Both the ERT and RF models were highly accurate in both training and validation stages. The results indicate that the ERT model is the best choice for predicting coagulant dosage in drinking water treatment plants and has the potential to improve operational efficiency and effectiveness in water treatment processes [7]. In summary, these studies demonstrate the effectiveness of using artificial neural network and random forest for predicting optimal dosage of coagulant. The use of Artificial Neural Network and Random Forest has been shown to improve the accuracy and efficiency of predicting the optimal coagulant dose, which can result optimizing the water treatment process resulting in cost savings and better overall performance of the treatment.

3. Methodology

3.1 Study Area and Dataset

This study focuses on the development of predictive models for determining optimal coagulant dosage in water treatment process, using data from the Parvati Water Treatment Plant, Pune, Maharashtra, India. The data covers period from 1st January 2022 to 10th November 2022, comprising 269 data points. The dataset includes essential raw water parameters such as pH, temperature, alkalinity, turbidity and conductivity, along with the coagulant dosage applied for the treatment. To determine the optimum dosage of coagulant two soft computing techniques have been used namely Random Forest and Artificial Neural Network. The model development and Testing methods are given in Model Formation (Section 3.2) below.

3.2 Model Formation

The Dataset received from Parvati Water Treatment Plant had a total of 9 water quality parameters: turbidity, color, hardness, calcium hardness, magnesium hardness, alkalinity, conductivity, and temperature for raw and treated water. In soft computing techniques, selection of input parameter is essential to mitigate the ‘curse of dimensionality’, a phenomenon leading to data sparsity and computational inefficiency with increasing dimensions. Selecting most important input parameters enhances model performance, interpretability, and efficiency by eliminating irrelevant features [8]. In this study based on domain knowledge and trial and error method, various input parameters have been selected and four models have been prepared with different input parameters. The description of the created models is given in Table 1. Below.

Table 1 Description of Models

Sr. No.	Models	Input Parameter	Output Parameter
1	Model A	Turbidity	Coagulant Dose
2	Model B	pH, Alkalinity, Temperature, Turbidity	Coagulant Dose
3	Model C	pH, Alkalinity, Temperature	Coagulant Dose
4	Model D	Alkalinity, Turbidity	Coagulant Dose

The models presented in Table 1, are trained with two data driven techniques named Random Forest and Artificial Neural Network. The present study deals with the model development in three phases as Phase 1, Phase 2 and Phase 3. These phases are according to the data arrangement for training and testing of models.

Phase 1: The important parameters has been identified (Table 1) for determining the coagulant dose as turbidity, pH, alkalinity and temperature. The data has been arranged on monthly basis and analyzed.

Phase 2: The data arrangement has been done based on the turbidity levels of the raw water. The dataset has been divided into training and testing sets ensuring that the training set included data with the highest and lowest turbidity values as well as a range of mid-range values. To improve models learning from the dataset.

Phase 3: The data has been organized in descending order based on the turbidity values. Since there has repetitions of raw turbidity values, a portion of 30% from each unique turbidity values allocated for testing purpose and 70% for training the models. The goal was to create a more balanced representation of data and increase the learning of the model by minimizing noise during the training process.

In this study models are developed by Random Forest using Weka (Version – 3.8.6) and Artificial Neural Network using MATLAB (R2023b). Developed models are evaluated using Scatter Plots, Root Mean Squared Error (RMSE) and Coefficient of Correlation (R).

The development and evaluation of Models has been explained in section 2.3 Random Forest and 2.4 Artificial Neural Network, below.

3.3 Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees during the training phase and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random Forest leverages the collective wisdom of numerous decision trees. Each tree contributes to the prediction process, and ensemble nature of Random Forest ensures robustness and accuracy.

In Random Forest, Bagging Percentage is the percentage of data given to one tree for the training purpose, after conducting some trials on varying bagging percentage on trial and error basis bagging percentage of 70% and 100% are decided for conducting trials. For testing of models, 3 techniques are used (i) Complete dataset for testing (ii) 30% dataset for testing and (iii) last 30% dataset for testing. The models are evaluated on the basis of Root Mean Squared Error (RMSE) and Coefficient of Correlation (R). The results obtained from these trials are given in Table 2 below.

Table 2 Random Forest Results

Model Name	Testing Method	Bagging Percentage			
		70%		100%	
		RMSE	R	RMSE	R
Model A	Complete Data	2.585	0.8137	2.5475	0.8196
	Random 30%	2.9229	0.7116	3.0435	0.6919
	Last 30%	4.3656	0.7278	4.3931	0.7156
Model B	Complete Data	1.1635	0.968	0.8781	0.9823
	Random 30%	1.9823	0.8733	1.9576	0.8765
	Last 30%	2.0992	0.8348	2.0695	0.8480
Model C	Complete Data	1.4003	0.9579	1.0340	0.9783
	Random 30%	2.9242	0.7107	3.0590	0.7035
	Last 30%	7.7102	-0.1155	7.8609	-0.1248
Model D	Complete Data	1.5236	0.9412	1.3331	0.9553
	Random 30%	2.1025	0.8545	2.1095	0.8536
	Last 30%	3.6653	0.7069	3.7310	0.6927

In Phase 1, after the training and testing of models, based on the results given in Table 2, on the basis of RMSE and R models are compared to each other to determine best performing models. Model B and Model C with RMSE value of 0.8781 and 1.0340 respectively and R value of 0.9823 and 0.9783 respectively found out to be the best performing models in Phase 1. The best performing models in Phase 1 of Random Forest are given in Table 3 below.

Table 3 Best Performing Models Random Forest

Model Name	Testing Method	Bagging Percentage	RMSE	R
Model A	Complete Data	100%	2.5475	0.8196
Model B	Complete Data	100%	0.8781	0.9823
Model C	Complete Data	100%	1.0340	0.9783
Model D	Complete Data	100%	1.3331	0.9553

In Phase 2, data arrangement for training of models is changed as discussed in Model Formation (Section 3.2). Results of the models in Phase 2 are given in Table 4 below. In Phase 2, after the training and testing of models a after comparison on the basis on RMSE and R based on the results given in Table 4 suggests, Model C and Model D with RMSE value 1.0660 and 1.2909 respectively and R value 0.9767 and 0.9583 respectively as best performing models. After comparison between Phase 1 and Phase 2 best performing models changed from Model B and C in Phase 1 to Model C and D in Phase 2. Phase 1 Models performed better with low RMSE values of Model B (0.8781) and Model C (1.0340) with higher R values 0.9823 and 0.9783 respectively.

Table 4 Random Forest Results

Model Name	Testing Method	Bagging Percentage			
		70%		100%	
		RMSE	R	RMSE	R
Model A	Complete Data	2.3460	0.8337	2.1007	0.8630
	Random 30%	5.5136	0.1324	5.7856	0.0549
	Last 30%	3.7661	0.5216	3.8233	0.5122
Model B	Complete Data	2.5752	0.8295	2.5573	0.8321
	Random 30%	3.6015	0.6588	3.7852	0.6273
	Last 30%	2.9248	0.7285	2.9079	0.7325
Model C	Complete Data	1.4273	0.9558	1.0660	0.9767
	Random 30%	3.1355	0.8017	2.9311	0.8272
	Last 30%	3.1588	0.7159	3.2040	0.7138
Model D	Complete Data	1.5002	0.9432	1.2909	0.9583
	Random 30%	2.7157	0.8422	2.8769	0.8212
	Last 30%	3.6981	0.5956	3.7702	0.5816

After all the trials in Phase 2, best trials for each model having the highest coefficient of correlation (R) and lowest Root Mean Squared Error (RMSE) are identified. To determine the best performing models in Phase 2 of Random Forest. In Table 5 below best performing models are given.

Table 5 Best Performing Models Random Forest

Model Name	Testing Method	Bagging Percentage	RMSE	R
Model A	Complete Data	100%	2.1007	0.8630
Model B	Complete Data	100%	2.5573	0.8321
Model C	Complete Data	100%	1.0660	0.9767
Model D	Complete Data	100%	1.2909	0.9583

In Phase 3, Models are again developed with different data arrangement having unique values of turbidity for both training and testing dataset to increase the learning of the model by giving non-repetitive values, and reduce the noise in the data as discussed in Model Formation (Section 3.2). To determine the best performing models a comparison on the basis of RMSE and R is done for results given in Table 6. After the comparison, the Model B and Model C are having the lowest RMSE values, suggesting a lower error in prediction and highest R values, suggesting that models predictions are nearly aligned to the actual dosage applied. Therefore, Model B and Model C with RMSE value 0.8757 and 1.0896 respectively with R value 0.9835 and 0.9791 respectively are found out to be best performing models in Phase 3, After Comparing Phase 3 results with Phase 1, for Model B the RMSE value reduced by 0.0024 and for Model C the RMSE values increased by 0.0556. As for Phase 3, Model C and Model D outperformed respective models in Phase 2 Model C and Model D with lower RMSE by 0.0236 and 0.0197 respectively and increased R values by 0.003 and 0.0132 giving a more powerful Model for Prediction of Optimum Dosage of Coagulant.

Table 6 Random Forest

Model Name	Testing Method	Bagging Percentage			
		70%		100%	
		RMSE	R	RMSE	R
Model A	Complete Data	2.1798	0.8777	2.1075	0.8861
	Random 30%	3.1602	0.6910	3.3627	0.6453
	Last 30%	2.2174	0.8575	2.4624	0.8386
Model B	Complete Data	1.1979	0.9684	0.8757	0.9835
	Random 30%	3.0305	0.7252	2.9458	0.7383
	Last 30%	1.6504	0.9367	1.6764	0.9407
Model C	Complete Data	1.5522	0.9552	1.0896	0.9791
	Random 30%	4.0701	0.4208	3.9370	0.4715
	Last 30%	4.9972	0.4686	5.2305	0.4238
Model D	Complete Data	1.3293	0.9574	1.0942	0.9715
	Random 30%	2.6026	0.8060	2.6676	0.7928
	Last 30%	1.8163	0.9420	2.2175	0.9173

After all the trials in Phase 3, to determine the best trials for each models a comparison on the basis or RMSE and R have been done based on the results shown in Table 6 above. The results of all the best performing trials for the respective models are given in Table 7 below.

Table 7 Best Performing Models Random Forest

Model Name	Testing Method	Bagging Percentage	RMSE	R
Model A	Complete Data	100%	2.1075	0.8861
Model B	Complete Data	100%	0.8757	0.9835
Model C	Complete Data	100%	1.0896	0.9791
Model D	Complete Data	100%	1.0942	0.9715

After Random Forest, Artificial Neural Network is used to develop the models for prediction of optimum dosage of coagulant. The development and evaluation of ANN models is discussed in Artificial Neural Network (Section 2.4) below.

3.4 Artificial Neural Network (ANN)

Artificial Neural Network is a computational model inspired by the structure and functionality of the human brain. It consists of interconnected nodes, or neurons, organized in layers: input layer, hidden layers and output layers. Each neuron receives input signals, process them using activation functions and transmits output signals to other neurons. ANN's flexibility and ability to varying input conditions make it powerful tool in soft computing for water treatment. Through iterative learning, ANN continuously improves its predictive performance, providing valuable insights for optimizing coagulant dosage in water treatment process. In Artificial Neural Network, the structure of the network consists of Input Layer, Hidden Layer and Output Layer. The nodes in Input Layers are equal to Input Parameters. The node in output layer is equal to the Output Parameter. The number of Hidden Layers and number of nodes in Hidden Layer is up to the researcher to decide. In this study ANN models are developed with different structures (number of layers and nodes) and trained using the Levenberg-Marquardt algorithm. The number of hidden layers and nodes are decided by trial and error for each Model. The results obtained from these trials are given in Table 8 below.

Table 8 Artificial Neural Network

Model Name	Structure	RMSE	R
Model A	1:2:1	0.295869	0.55906
	1:3:1	0.296973	0.59831
	1:4:1	0.295847	0.55648
	1:10:1	0.296975	0.60150
	1:1:1	0.295778	0.55204
	1:30:1	0.303422	0.60251
	1:15:1	0.296973	0.60161
	1:7:1	0.296973	0.60239
Model B	4:2:1	0.197526	0.73867
	4:3:1	0.326720	0.62387
	4:4:1	0.032494	0.67305
	4:8:1	0.555120	0.71093
	4:1:1	0.104518	0.64032
	4:6:1	0.121200	0.40730
	4:20:1	0.553156	0.04779
	4:5:1	0.037323	0.57510
Model C	3:2:1	0.892169	0.68263
	3:3:1	0.019346	0.37977
	3:1:1	0.162775	0.14005
	3:6:1	0.20081	0.15849
	3:4:1	0.007028	0.3823
	3:5:1	0.090143	0.34826
	3:12:1	0.449872	0.16496
	3:30:1	6.894182	0.031395
Model D	2:1:1	0.014714	0.65072
	2:2:1	0.338919	0.36554
	2:4:1	0.339566	0.36203
	2:20:1	0.267904	0.35131
	2:3:1	0.01863	0.30053
	2:5:1	0.23762	0.42168
	2:6:1	0.092572	0.50268
	2:8:1	0.329243	0.41742

In Phase 1, after all the trials, to find the best performing models a comparison on the basis of RMSE and R have been done using results shown in Table 8 above. Based on the results a comparison between all the trials for respective model have been done to identify the best trials having low RMSE and high R values. Best trials for each model are given in Table 9 below. On the basis of best models in Table 10, Model C and Model D with low RMSE values of 0.007028 and 0.014714 respectively and R values of 0.38230 and 0.65075 respectively are the best performing models in Phase 1 of Artificial Neural Network. In Phase 2, data arrangement is changed as discussed in Model Formation (Section 3.2). The results of the developed models are given in Table 11 below.

Table 9 Best Performing Models Artificial Neural Network

Model Name	Structure	RMSE	R
Model A	1:1:1	0.295778	0.55204
Model B	4:4:1	0.032494	0.67305
Model C	3:4:1	0.007028	0.38230
Model D	2:1:1	0.014714	0.65072

Table 10 Artificial Neural Network

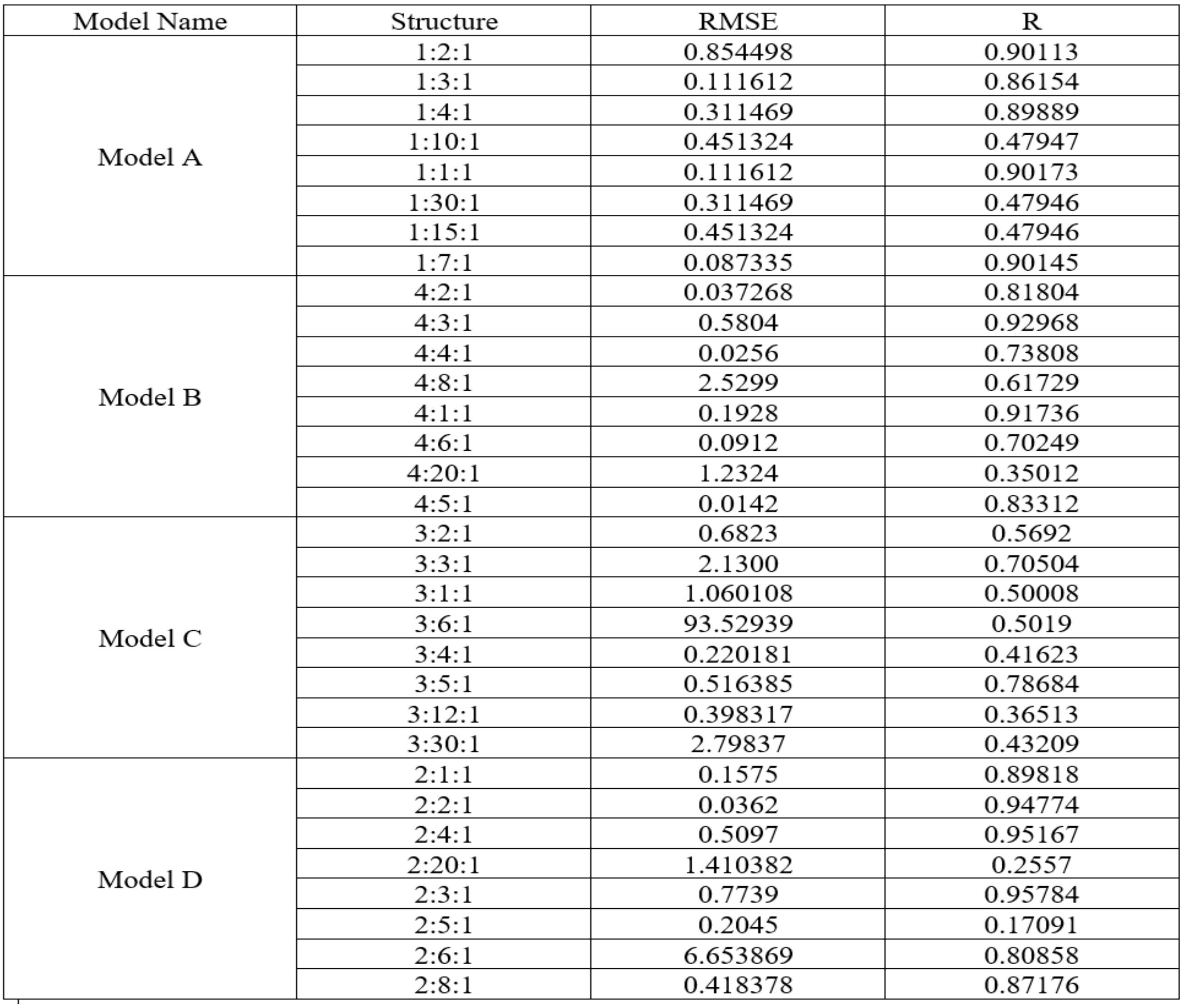
Model Name	Structure	RMSE	R
Model A	1:2:1	0.887305	0.48787
	1:3:1	0.887304	0.48894
	1:4:1	0.917228	0.49563
	1:10:1	0.887307	0.47865
	1:1:1	0.919719	0.48861
	1:30:1	0.830593	0.51066
	1:15:1	0.887307	0.47905
	1:7:1	0.887303	0.48842
Model B	4:2:1	0.586928	0.64023
	4:3:1	0.70039	0.77764
	4:4:1	0.681187	0.72449
	4:8:1	0.586552	0.78692
	4:1:1	0.558899	0.55687
	4:6:1	0.881994	0.69216
	4:20:1	1.711464	0.34501
	4:5:1	0.574329	0.76939
Model C	3:2:1	1.04763	0.65566
	3:3:1	1.120785	0.65497
	3:1:1	0.752372	0.51564
	3:6:1	1.301879	0.63331
	3:4:1	1.143225	0.66004
	3:5:1	1.295553	0.64136
	3:12:1	0.064359	0.41226
	3:30:1	0.735188	0.12194
Model D	2:1:1	0.968655	0.53436
	2:2:1	0.931502	0.68097
	2:4:1	0.940856	0.64172
	2:20:1	0.528337	0.58993
	2:3:1	0.937208	0.67925
	2:5:1	0.941622	0.59655
	2:6:1	1.03782	0.5917
	2:8:1	0.943158	0.56786

In Phase 2, based on results given in Table 11 comparison on the basis on RMSE and R is done to determine best performing model in Phase 2. Model C and Model D with RMSE value of 0.064359 and 0.528337 respectively and R value of 0.41226 and 0.58993 respectively are the best performing models. After comparison between Phase 1 and Phase 2, In Phase 2 Model B slightly improved its R by 0.02996. The results of best performing models in Phase 2 are given in Table 12 below. In Phase 3, Models are again trained with different data arrangement with unique values of turbidity as discussed in Model Formation (Section 3.2). The results of trials conducted are given in Table 13 below.

Table 11 Best Performing Models Artificial Neural Network

Model Name	Structure	RMSE	R
Model A	1:30:1	0.830593	0.51066
Model B	4:1:1	0.558899	0.55687
Model C	3:12:1	0.064359	0.41226
Model D	2:20:1	0.528337	0.58993

Table 12 Artificial Neural Network



Model Name	Structure	RMSE	R
Model A	1:2:1	0.854498	0.90113
	1:3:1	0.111612	0.86154
	1:4:1	0.311469	0.89889
	1:10:1	0.451324	0.47947
	1:1:1	0.111612	0.90173
	1:30:1	0.311469	0.47946
	1:15:1	0.451324	0.47946
	1:7:1	0.087335	0.90145
Model B	4:2:1	0.037268	0.81804
	4:3:1	0.5804	0.92968
	4:4:1	0.0256	0.73808
	4:8:1	2.5299	0.61729
	4:1:1	0.1928	0.91736
	4:6:1	0.0912	0.70249
	4:20:1	1.2324	0.35012
	4:5:1	0.0142	0.83312
Model C	3:2:1	0.6823	0.5692
	3:3:1	2.1300	0.70504
	3:1:1	1.060108	0.50008
	3:6:1	93.52939	0.5019
	3:4:1	0.220181	0.41623
	3:5:1	0.516385	0.78684
	3:12:1	0.398317	0.36513
	3:30:1	2.79837	0.43209
Model D	2:1:1	0.1575	0.89818
	2:2:1	0.0362	0.94774
	2:4:1	0.5097	0.95167
	2:20:1	1.410382	0.2557
	2:3:1	0.7739	0.95784
	2:5:1	0.2045	0.17091
	2:6:1	6.653869	0.80858
	2:8:1	0.418378	0.87176

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1408 - 1420
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0194

To find the best performing models a comparison between RMSE and R have been done from results given in Table 13. Model B and Model D with RMSE value of 0.0142 and 0.0362 respectively with R value of 0.83312 and 0.94774 respectively, found out as best performing models in Phase 3, Comparing it with Phase 1 Model B and Model D, the R values increased by 0.16007 and 0.29702 respectively. RMSE values are slightly increased in Phase 3 compared to Phase 1. By comparing Phase 2 with Phase 3, for Model B and Model D RMSE values decreased by 0.544699 and 0.492137 respectively and R values increased by 0.27625 and 0.35781 respectively in Phase 3. The results of best performing models in Artificial Neural Network Phase 3 are given in Table 14 below. After all the models have been developed, to compare both the techniques and models, a comparison on the basis of RMSE and R has been done.

Table 13 Best Performing Models Artificial Neural Network

Model Name	Structure	RMSE	R
Model A	1:7:1	0.087335	0.90145
Model B	4:5:1	0.0142	0.83312
Model C	3:4:1	0.220181	0.41623
Model D	2:2:1	0.0362	0.94774

Table 15 shows the best performing models in both techniques used namely (i). Random Forest and (ii). Artificial Neural Network. The comparison has been done for final models developed in Phase 3 and discussed in Results and Discussion (Section 3). The summary of all best performing models and techniques are given in Table 15 below.

Table 14 Best Performing Models and Techniques

Technique	Phase					
	1		2		3	
	RMSE	R	RMSE	R	RMSE	R
Random Forest	0.8781	0.9823	2.5573	0.8321	0.8757	0.9835
	1.0340	0.9783	1.0660	0.9767	1.0896	0.9791
Artificial Neural Network	0.0147	0.6507	0.5283	0.5899	0.0362	0.9477
	0.0324	0.6730	0.5588	0.5568	0.0142	0.8331

After the development of models by Random Forest and Artificial Neural Network, further analysis to determine best technique and input parameters for the prediction of optimum dosage of coagulant using scatter plots is given in Results and Discussion (Section 3) below.

4. Results and Discussion

4.1 Random Forest

Model A

The RMSE value of 0.8757 implies a low prediction error, indicating that, typically, the model's predictions deviate by 0.8757 PPM from the actual coagulant dose.

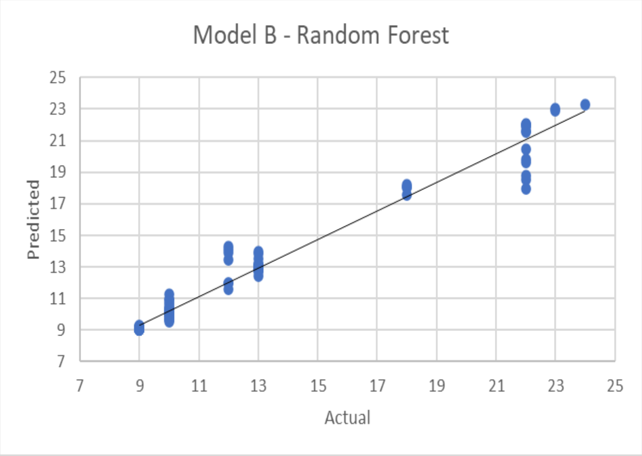


Figure 1 Scatter Plot Random Forest Model A

Correlation Coefficient (0.9835) indicated a strong positive linear relationship between the predicted and actual dose. Indicating that the model's predictions are closely associated with the actual dose.

Model B

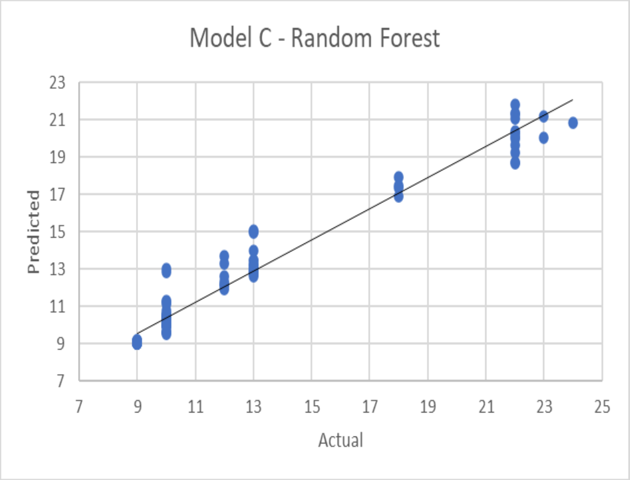


Figure 2 Scatter Plot Random Forest Model B

The RMSE value of 1.0896 implies a low prediction error, indicating that, typically, the model's predictions deviate by 1.0896 PPM from the actual coagulant dose. Correlation Coefficient (0.9791) indicated a strong positive linear relationship figure 2 between the predicted and actual dose. The data points are close to the line of best fit. Indicating that the model's predictions are closely associated with the actual dose. Comparing between Model B and Model C of Random Forest, Model B performs better with a low RMSE value of 0.8757 and a higher coefficient of correlation 0.9835. The model gives strong positive linear relationship between actual and predicted values also by observation of Figure 1 the data points are close to the best fit line showcasing good prediction capabilities of the model.

4.2 Artificial Neural Network

Model A

The RMSE value of 0.0142 implies a low prediction error, indicating that, typically, the model's predictions deviate by 0.0142 PPM from the actual coagulant dose.

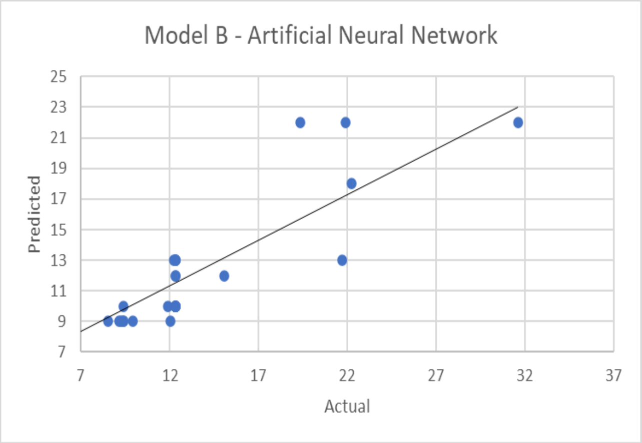


Figure 3 Scatter Plot Artificial Neural Network Model A

The model has moderate Correlation Coefficient (0.83312) and follows a positive linear relationship between the predicted and actual dose. Indicating that the model's predictions are closely associated with the actual dose. As the data points are not close to the best fit line and are scattered away from the best fit line the model results in a moderate coefficient of correlation of 0.83312. Having low RMSE value of 0.0142 Artificial Neural Network Model B outperforms Random Forest Model B and Model C giving better predictions with low prediction error.

Model B

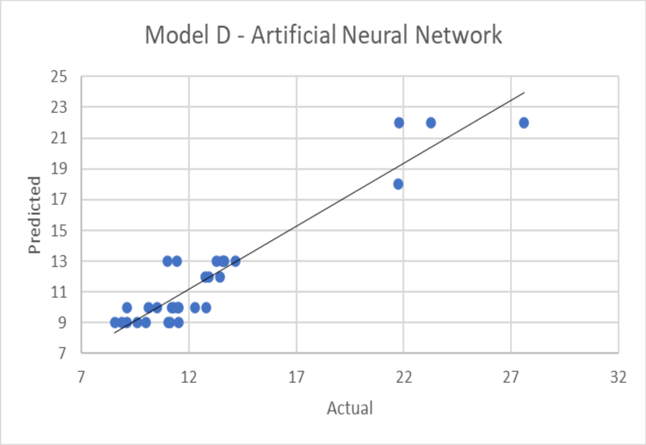


Figure 4 Scatter Plot Artificial Neural Network Model D

The RMSE value of 0.0362 implies a low prediction error, indicating that, typically, the model's predictions deviate by 0.0362 PPM from the actual coagulant dose. The model has high Correlation Coefficient (0.9477) indicated a strong positive linear relationship between the predicted and actual dose. The data points are close to the line of best fit. Indicating that the model's predictions are closely associated with the actual dose. Comparing between Model D and Model B of ANN, ANN Model B's RMSE is 0.022 lower than ANN Model D and outperforms all the other Random Forest and Artificial Neural Network models. As the Difference between RMSE of ANN Model B and Model D is 0.022, both models are performing good in terms of RMSE. Comparing on the basis of scatter plots Models D's predictions are closer to the best fit line (Figure 4), as for Model B's predictions are scattered around the best fit line (Figure 3). Comparing on the basis of predictions Model D's predictions are more aligned with the actual dose giving Model D a higher coefficient of correlation. In conclusion, even though Model B is having a low RMSE, Model D predicts coagulant dose better. Making Model D as the best performing model and ANN as the best performing technique for the prediction of optimum dosage of coagulant.

Conclusion

In this study a comparison between two soft computing techniques: Random Forest and Artificial Neural Network has been done for the prediction of optimum dosage of coagulant in water treatment plant. 4 Models with different input parameters and optimum dosage of coagulant as output parameter have been tested for determination of best performing Model and Technique for the prediction of Optimum Dosage of Coagulant. In Random Forest Model B and Model C performed best with RMSE value of 0.8757 and 1.0896 respectively and a high coefficient of correlation R value of 0.9835 and 0.9791 respectively. Random Forest Model B with a 0.2139 PPM lower RMSE value than Model C, proved to be better performing model for prediction of coagulant dosage. Lastly In Artificial Neural Network Model D and Model B performed best with RMSE 0.0362 and 0.0142 respectively and a high Coefficient of Correlation (R) value of 0.9477 and 0.83312 respectively. ANN Model B's RMSE is 0.022 lower than ANN Model D and outperforms all other models in Random Forest and Artificial Neural Network. In this comprehensive analysis, Artificial Neural Network Model D (Input Parameters - Alkalinity, Turbidity) and Model B (Input Parameters – pH, Alkalinity, Temperature, Turbidity) emerged as standout performers showcasing remarkable accuracy with low Root Mean Squared Error (RMSE) values of 0.0362 and 0.0142 respectively, complemented by high correlation coefficients (R) of 0.9477 and 0.83312 respectively. Making Model D and Model B the best performing model and ANN as the best performing technique for the prediction of optimum dosage of coagulant.

References

[1]. Mohamed Hassen Baouab, Semia Cherif (2018). Prediction of the optimal dose of coagulant for various potable water treatment processes through artificial neural network. Journal of Hydroinformatics, 20(6), 942-952. doi: 10.2166/hydro.2018.014.

[2]. Ruba Alsaeed, Bassam Alaji, Mazen Ebrahim. (2021). Predicting Turbidity and Aluminium in Drinking Water Treatment Plants Using Hybrid Network (GA-ANN) and GEP. Drinking Water Engineering and Science Discussion. doi: 10.5194/dwes-2021-8

[3]. Alka Kote, Dnyaneshwar Wadkar. (2019). Modelling of Chlorine and Coagulant Dose in a Water Treatment Plant by Artificial Neural Networks. Engineering, Technology and Applied Science Research, 9(3), 4176-4181. doi: 10.48084/etasr.2725.

[4]. Ayse B. Sengul & Zeliha Gormez. (2013). Prediction of Optimal Coagulant Dosage in Drinking Water Treatment by Artificial Neural Network. 1st International EwaS-

MED International Conference

[5]. Mohammed Achite, Saeed Samadianfard, Nehal Elshaboury, Milad Sharafi. (2023). Modelling and Optimization of Coagulant Dosage in Water Treatment Plants using Hybridized Random Forest Model with Genetic Algorithm Optimization. Environmental, Development and Sustainability, Vol 25, 11189-11207. doi: 10.1007/s10668-022-02523-z

[6]. Dongsheng Wang, Le Chen, Taiyang Li, Xiao Chang, Kaiwei Ma, Weihong You, Chaoqun Tan. (2023). Successful Prediction for Coagulant Dosage and Effluent Turbidity of a Coagulation Process in a Drinking Water Treatment Plant Based on The Elman Neural Network and Random Forest Models. Environmental Science: Water Research & Technology. doi:10.1039/D3EW00181D

[7]. Salim Heddam. (2021). Extremely Randomized Tree: A New Machines Learning Method for Predicting Coagulant Dosage in Drinking Water Treatment Plant. Water Engineering Modeling and Mathematic Tools. 475-489. doi: 10.1016/B978-0-12-820644-7.00013-X.

[8]. Isabelle Guyon & Andre Elisseeff. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 1157-1182. doi: 10.1162/153244303322753616.