Journal of Soft Computing in Civil Engineering 7-4 (2023) 110-131

Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Soft Computing Techniques for Predicting Chemical Oxygen Demand in River Water

Pali Sahu[1*], Shreenivas N Londhe[2], Preeti S Kulkarni[3]

1. Research Scholar, Civil Department, Vishwakarma Institute of Information Technology (VIIT), Pune, India
2. Professor, Civil Department, Vishwakarma Institute of Information Technology (VIIT), Pune, India
3. Associate Professor, Civil Department, Vishwakarma Institute of Information Technology (VIIT), Pune, India
Corresponding author: *palisahu18@gmail.com*

https://doi.org/10.22115/SCCE.2023.366329.1544

ARTICLE INFO

Article history:
Received: 19 October 2022
Revised: 22 April 2023
Accepted: 23 April 2023

Keywords:
ANN;
MGGP;
Soft computing;
Modelling;
Water quality;
Chemical oxygen demand.

ABSTRACT

Organic matter in water is assessed through Chemical oxygen demand (COD). COD prediction utilizing Data driven technique (DDT) has shown to be promising and may be utilized as supplemental techniques due to the time-consuming procedure and nonlinear correlations between the factors. The current study aims to determine how well three different DDT, namely Artificial Neural Network (ANN), Multi-Gene Genetic Programming (MGGP), and Model Tree (M5T), can estimate the concentration of COD in water taken from three different sections of the Mula, Mutha, and Mula-Mutha Rivers in Pune, India. The performance of the models demonstrates that both ANN and MGGP worked brilliantly, with a correlation coefficient (between observed and projected values) that was more than 0.88 and a root mean square value of 0.7 mg/l across all three parts. The input frequency distribution in MGGP and the input variable coefficient in M5T indicate that both techniques can identify the influential factors. MGGP and MT score with readily available equations as model.

How to cite this article: Sahu P, Londhe SN, Kulkarni PS. Soft computing techniques for predicting chemical oxygen demand in river water. J Soft Comput Civ Eng 2023;7(4):110–131. https://doi.org/10.22115/scce.2023.366329.1544

1. Introduction

River water quality is mainly assessed on the basis of organic matter content, like biochemical oxygen demand (BOD) and chemical oxygen demand (COD). BOD is the quantity of dissolved oxygen consumed by aerobic microorganisms to oxidize organic matters [1]. The COD measures the total amount of oxygen-consuming compounds in the full chemical breakdown of organic molecules in water. It is a key measure for monitoring water quality and defining the sort of organic load present [2]. The traditional approach for COD analysis is done by multistep chemical technique i.e., reflux method, which is a nearly 2-hour operation, in which sealed tubes filled with chemicals are heated to 150 degrees, creating tremendous pressure beneath the tubes, thus making the analysis risky due to the risk of explosion [3,4]. Additionally, the presence of various inorganic interfering matter, most of the time distorts the outcome of the analysis and makes this process less accurate [3]. Hence, with limited resources and time, cost-effective model development with accurate modelling and forecasting of river water quality is necessary for ecologically sound water management. Traditional deterministic and probabilistic models need precisely established rate coefficients, many of which are temporally and geographically unique, for a variety of hydro-chemical, physical, and biological processes [4]. To circumvent these constraints, researchers have developed Data-driven techniques (DDT) to model water quality parameters. DDT is a computational approach that uses system state variables (input-output) to replace physical-behaviour-based knowledge-driven models [5]. DDT can be Soft computing and Hard computing. Soft computing employs partial truth, ambiguity, and approximation. Soft computing has grown in popularity due to its various features such as optimisation, intelligent control, decision-making, and nonlinear programming [5]. Adaptive-network-based fuzzy inference system (ANFIS), artificial neural network (ANN), Multi-Layer-Perceptron (MLP), genetic algorithm (GA), and Fuzzy Logic (FL) are some of the most often utilized DDT for water quality metrics [5–10]. Over the last decade, several articles have asserted that DDT based models accurately replicate dissolved oxygen (DO), BOD concentration, and other important water quality variables.

Out of the various techniques mentioned above, ANNs have been utilised for many water and environmental research like Mehr [11] employed GP-SARIMA to improve long-term streamflow forecasting in a lake-river system of Oulujoki River, Finland. A model for one-step-ahead streamflow predictions is tested. The results demonstrated that a combination of correlogram and average mutual information (AMI) analysis may help to pick the best lags for streamflow model predictors. Karami [12] employed Neural Network based method (NN) to model and simulate rate of evaporation for Garmsar city of Iran. Testing phase results shows that the NN model is able to simulate evaporation with minimum error. While Palani et al. [6] utilised ANN to predict and forecast temperature, salinity and DO in Singapore coastal waters. Singh et al. [7] utilized ANN technique to simulate the DO and BOD levels in the Gomti River (India). `Najah et al. [13] and Basant et al.[14] employed linear and nonlinear modelling using ANN's feed-forward-back-propagation and partial least squares (PLS) regression to forecast the DO and BOD levels in river water at the same time. Results revealed that both models could predict DO and BOD levels, but the non-linear (ANN) model outperformed the linear (PLS)

model. The application of ANN for the prediction and modelling of COD removal from antibiotic aqueous system by the Fenton process was explored by Elmolla et al. [15] and found that the outcome was extremely close to the actual data, with high value of R^2 (coefficient of determination) i.e., 0.997 and an MSE of 0.000376. Emamgholizadeh et al. [16] employed MLP, radial basis network (RBN) and ANFIS to model three important water quality parameters (DO, BOD, and COD) for highest water flow Iranian Karoon River. Results demonstrated that estimated values of DO, BOD, and COD by using both ANN and ANFIS were reflecting balanced scatter with observed values (R= 0.86, 0.94, 0.96 respectively). Akilandeswari and Kavitha [8] estimated COD concentration for textile effluent using the ANFIS and multiple linear regressions (MLR). As a consequence, the ANFIS technique outperformed the MLR in modelling COD concentration. Convolutional Neural Network-Long Short-Term Memory Network (CNN-LSTM) technology, which is based on an attention mechanism, was used by Xijuan et al. [17] to anticipate the water quality of the Yellow River in China. The results show that the hybrid model performs better than the conventional NN model in solving nonlinear time series prediction issues. In Kayseri, Turkey, Ozkan [18] used MLP to forecast BOD in a sewage water treatment facility using temperature, suspended solids (SS), COD, total dissolved solids (TDS), total nitrogen, and total phosphorous. The correlation value was 0.915, and the ANN technique for modelling DO prediction was determined to be satisfactory. Dogan et al. [19] used chemical oxygen demand, flow (Q), temperature, nitrites, total ammonia, and nitrate variables to explore the possibility of an `MLP model to increase the precision of BOD prediction in the Melen River (Turkey) and findings were satisfactory (R=0.89). Heddam & Kisi [9], explored different approach i.e., multi-variate adaptive regression spline (MARS), model tree (MT) in order to model DO and evaluate its correctness using a least-squares support vector machine (LSSVM). Compared to LSSVM, MARS fared the best, with a R of 0.965 and RMSE and MAE of 0.547mg/L and 0.386 mg/L, respectively. Similarly, Mehr et al. [20] used a multi-step evolutionary search algorithm in which high-performance rain-borne genes from a multigene GP solution is combined through a classic Genetic Programming engine for predicting 1 month ahead rainfall measurements from meteorology stations in Lake Urmia Basin, Iran. The model proposed outperforms the benchmark models: standard GP and autoregressive state-space in both the rain gauge stations. The results show around 24% and 60% improvement (average of two case studies) in terms of Root mean square error and Nash-Sutcliffe coefficient of efficiency metrics. The proposed model: Multiple genetic programming (MGP) eliminates genes of lower performance and only allows those of higher performance to contribute to the final solution is stated as the reason for better accuracy.

The majority of research reveals that most of the work is focused on the prediction of DO and BOD for river water rather than COD and employs ANN, ANFIS, and a small number of studies also used LGP and SVM techniques to simulate water quality matrices. Despite the fact that these models do enhance accuracy to some degree, they also have certain drawbacks. For instance, ANN's major operational hurdle is its inability to generate the final outcome using a simple mathematical equation, making it less portable, whereas ANFIS was limited by its computing complexity. As a result, the established model for monitoring water quality has to be

revised in order to reduce the inaccuracies, computational complexity, and over-fitting issues that plagued prior approaches. In light of the above information, to the author's knowledge, no research has been conducted to leverage the potential of data-driven methodologies such as multi-gene genetic programming (MGGP) and model tree (M5T) for the prediction of water quality variables. Thus, current study is focused to utilize Multi gene genetic programming and Model tree with M5 algorithm to predict COD content for Mula-Mutha river Pune, India and compare the results with ANN model [8,15,18,19].

The water quality data for Mula, Mutha and Mula-Mutha River were employed over the last fifteen years (2003-2018) to attain this goal. The outcome of three developed models were analyzed and compared using root mean square error (RMSE), mean absolute relative error (MARE), and coefficient of correlation (R) statistical error measures, as well as visual presentation of values on graphs and scatter plots between actual recorded and model-predicted values. The next part will discuss the techniques employed, followed by information on the water quality data and study area. The results and discussions will follow next and the study will be concluded by conclusion.

2. Techniques utilized

Soft computing techniques treat human brain as their role model and mimic the ability of the human mind to effectively employ modes of reasoning that are approximate rather than exact. They do not assume any mathematical model a priori and hence are more flexible in data mining. The underlying idea of soft computing is to provide tractability, resilience, and a better connection with reality by allowing for imperfection, ambiguity, and partial truth. In the current study soft computing techniques like Artificial Neural Network and Genetic Programming are utilized along with other Model Tree with M5 technique.

2.1. Artificial neural network

Artificial Neural Network is a massively parallel, distributed processor with a natural tendency for storing experiential information to make it accessible for usage. In two aspects, it resembles the human brain: it utilizes information to recognize complex nonlinear behavior or patterns acquired by the network throughout the learning process, and it retains knowledge using synaptic weights, which are the strengths of interneuron connections. In general, an ANN is made up of three layers: an input layer, one or more hidden layer(s), and an output layer. Synaptic weights, biases, and transfer functions link the hidden layer to the other layers. Error functions are based on network output against aim. Modifying weights and biases backward using algorithms reduces error until the desired output accuracy is obtained. Training ANN via feed forward back propagation distributes error backward and processes weight and bias forward. Until output accuracy is achieved, the training cycle is repeated. The network's weights and biases can be used to validate unknown data after training. The connection between weights, input, and output is shown in Figure 1. Readers may refer to [5,10,21–24] for more information on how an ANN works in detail.

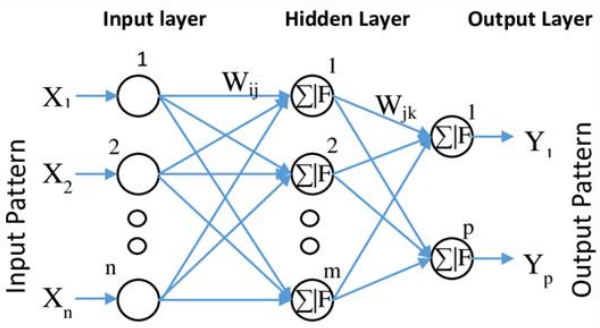


Fig. 1. Typical three-layer network architecture of ANN [25].

2.2. Model tree

Model Tree employs a divide-and-conquer strategy to offer rules for a linear model to arrive at the leaf node. The linear models are then used to determine how much each parameter contributes to the total anticipated value. Quinlan's M5 technique is reconstructed as M5 Model Tree (M5T) to trigger decision/regression trees of models [26]. M5T is a decision tree that incorporate a traditional decision tree with the possibility to use linear regression algorithms at the last nodes. To begin, a tree is built using the decision tree technique, and the splitting approach is utilized to curtail the standard deviation (sd) in the intra-subset of the M5 Model tree, resulting in linear models in the leaf node [27,28]. The M5T approach utilizes the two important steps: the tree growth step (splitting) and the tree pruning step [29,30]. Figure 2 shows how the M5 method of Model tree splits the input X_1 X_2 input variables into multiple alternative linear regression functions as model namely LM1-LM6 at leaf node. The model equation is $y = b_0 + b_1X_1 + b_2X_2$, where b_0, b_1, and b_2 are linear regression constants and depicts the relationship between branches as a tree diagram [29,31] are references for MT readers.

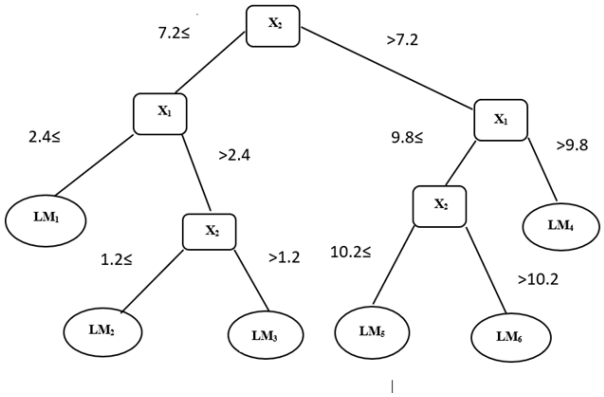
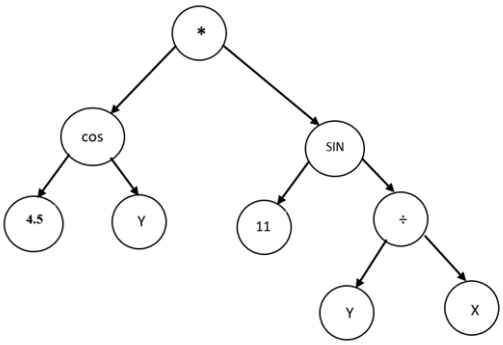


Fig. 2. Diagram of model tree with LM 1-6 at the leaves.

2.3. Genetic programming

Genetic programming (GP) is a domain-agnostic approach for resolving problems by breeding a population of randomly produced computer programs genetically. It belongs to the category of supervised machine learning, in which the answer is found in a program space rather than a data space. Traditional GP solutions are expressed as tree structures and stated in a functional programming language. Genetic Programming (GP) is based on the survival of the fittest

concept, which states that the fittest will live and participate in the next generation's evolution by breeding. The three genetic operations for breeding are as follows: Reproduction, Mutation, and Crossover.

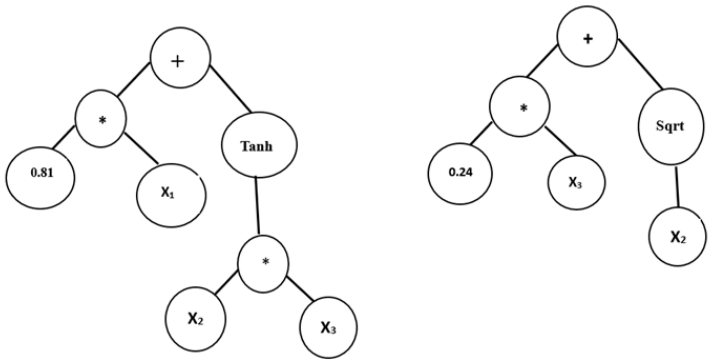


$$(4.5\ COS\ Y) * (11\ SIN\ (\frac{Y}{X}))$$

Fig. 3. Example of standard genetic programming (GP) tree representation for the function.

A few GP variants include GEP, LGP and MGGP. GEP employs defined length linear genomes, also known as chromosomes, to describe computer programmes in the form of expression trees. The chromosome is made up of a linear, symbolic string of single or multiple genes that is fixed in length. The ability of GEP chromosomes to code ETs of all sizes and forms will also be demonstrated, despite their constant length. A specific type of linear representation is used for computer programmes in LGP, a linear version of GP. Instead of functional genetic programmes that are restricted to a single linear list of nodes, the term "linear" refers to the shape of the (imperative) programme representation. Expressions from functional programming languages, such as LISP, are swapped out for imperative programming language programmes, like C or C++, in LGP. LGP's main characteristics are the graph-based data flow from indexed variables and programming in a low-level language that allows solutions to be directly changed into binary machine codes and run without an interpreter [32]. It is intended to build "multi-gene" mathematical models for specific response (output) data. In general, nonlinear input variables are merged into a low-order linear weighted tree (by limiting the gene or tree depth). The evaluation of a single tree (model) expression is used in the standard GP form. Multigene individuals in MGGP are made up of several genes, each of which corresponds to the "conventional" GP tree expression [32,33]. For MGGP model development firstly population and generation size are determined by the issue's complexity and the number of possible solutions. To construct models with the least level of uncertainty, several populations and generations are studied. Genes are changed by natural selection via mutations and crossings to produce offspring. The mutation mechanism selects branches and sub-nodes, replacing each with a random subtree. During the crossover procedure, random parent tree terminals or branched nodes are picked and swapped. This procedure is repeated until the termination requirement is reached, improving model fitness. Although a maximum number of genes (Gmax) of an individual and maximum depth of tree (Dmax), directly influence the size of the search space and the number of solutions explored within it. The MGGP algorithm's success frequently increases with these settings. Finally, the best model is chosen from the output values based on simplicity and fitness. Parameters allow

the user to change the model's simplicity (e.g., Gmax or Dmax). Figure 4 depicts a typical MGGP model built from two standard GP trees. With input variables x_1, x_2, and x_3, this model predicts an output variable. Although linear in the parameters concerning the coefficients variables d_0, d_1& d_2, the given model structure incorporates nonlinear components (e.g., tan, sin, sqrt. etc.). The fundamental arithmetic operators (/, x, +, - etc.) and Boolean logic functions are included in the function set (sin, cos, tanh, etc.). For details of GP, readers are referred to [33,34] and MGGP readers are referred to [34].



$$d_0 + d_1(0.81x_1 + Tanh(x_2x_3) + d_2(0.24x_3 + Sqrt(x_2))$$

Fig. 4. Example of MGGP symbolic methods.

3. Study area

The current research is focused on Pune city, a developing city in Maharashtra, which is blessed with the rivers Mula and Mutha and are major sources of water supply. Pune is located at 18° 31' 22.45" of North and 73° 52' 32.69" of East, near the western boundary of the Deccan Plateau. Pune lies on the downwind side of the Sahyadri hills and Western Ghats, 1837 feet above the sea level, at the conflux of the Mula and Mutha rivers, which are tributaries of the River Bhima. The Mula River flows roughly 64 kilometers from its source in the Pune District's hilly areas, with 40 kilometers of it being mountainous terrain. It then approaches Pune from the north-west, passing through thickly populated districts before meeting the Mutha River [35]. Within the Pune Metropolitan Region, many small town and small & micro scale (SSI) industries like paper-pulp mills and sugar mills, including some agriculture runoff, are responsible for garbage generation in the Mula River [36].

The Mutha River rises from Western Ghats and flows roughly 15 kilometers eastward until merging into the Mula River near Pune [37]. Along the Mutha River in the Pune Metro-Politian Region, there are several villages and historic city areas.

After the confluence of the Mutha and Mula Rivers at Sangamwadi, the united Mula-Mutha River runs through the city of Pune (Fig. 5) and then travels downstream to merge with Bhima River, a significant tributary of the Krishna River going southeast. The Mula-Mutha River is the most contaminated since it carries trash from sewage treatment plants and common effluent treatment plants [38]. The data was collected from stations M1, M2 and M3 as shown in figure 5 below.

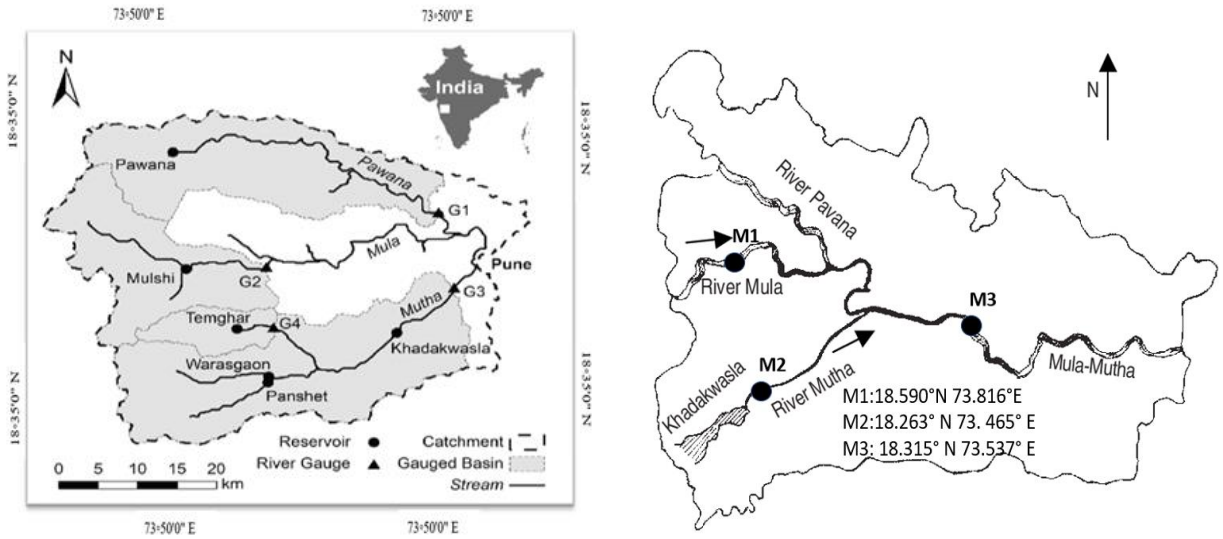


Fig. 5. Detailed location map of water study area Mula-Mutha River (www.mahap.gov.in).

4. Data set

Monthly water quality data for the rivers: Mula, Mutha and Mula-Mutha, requisite for the present research was received from Hydro Nashik Maharashtra, India for the years 2003 to 2018.The statistical analysis of the data from the year 2003 to 2018 for the reaches of the Mula, Mutha, and Mula-Mutha rivers utilized in this research is provided in table 2, Statistical analysis provides insight into or underlying pattern of any data, according to table 2, the skewness coefficient (Csx) is somehow very less for the majority of data sets, which is considered desirable since a high value of skewness has a detrimental effect on ANN performance. Additionally, coefficients of variation indicated significant changes in the Mula-Mutha River as compare to Mula River and Mutha River, owing to the river's route through various townships and the presence of numerous untreated wastewaters drains and tributaries.

Table 1
Correlation factor of COD with various influencing parameters.

Station	Mutha	Mula	Mula-Mutha
Variables	COD	COD	COD
BOD	0.97	0.85	0.88
DO	-0.88	-0.73	-0.46
EC	0.77	0.7	0.005
Hardness	0.25	0.66	0.112
pH	-0.31	-0.21	-0.579
TS	0.73	0.69	-0.26
TDS	0.78	0.71	-0.145
No_3-No_2	0.051	0.078	-0.458
No_3-N	0.089	0.338	-0.019

Table 2
Statistical analysis of entire hydro-chemical data for Mutha, Mula and Mula-Mutha station.

Station	Parameters	Unit	Min	Max	Mean	Sd	Csx	CV%
Mutha River	COD	mg/L	1.6	140	22.65	23.52	1.35	103.8
	DO	mg/L	0.2	9.2	5.7	3.1	1.24	54.3
	BOD	mg/L	0.4	38	0.8	9.5	1.54	118.75
	EC	µs/cm	2	640	216	172	1.66	79.62
	TS	mg/ L	50	413	176	98	0.62	55.68
	Hardness	mg/ L	24	231	88.76	54.14	1.15	60.95
	TDS	mg/ L	13	381	134	102	1.59	76.1
	pH	Unit	6.7	8.9	7.9	0.5	0.48	6.32
	No_3-N	mg/ L	0.01	11.31	0.24	0.86	1.04	358
	No_3-No_2	mg/ L	0.01	0.35	0.26	0.45	0.07	173
Mula River	COD	mg/L	1.6	76.8	21.19	14.6	1.39	66.6
	DO	mg/ L	0.28	9	4.6	2.7	1.23	58.5
	BOD	mg/ L	0.8	38	3.2	6.8	1.54	212
	EC	µs/cm	62	712	304	192	1.17	9.5
	TS	mg/ L	55	471	211	109	1.06	51.6
	TDS	mg/ L	45	434	180	108	1.05	6
	Hardness	mg/ L	24.6	284	102.3	26.5	0.53	25.9
	pH	Unit	6.93	8.75	7.7	0.3	0.08	3.8
	No_3-N	mg/L	0.01	10.6	0.29	1.29	1.15	445
	No_3-No_2	mg/ L	0.02	3.88	0.27	0.53	0.96	196
Mutha-Mula River	COD	mg/L	5.6	162	29.62	18.99	1.14	64.11
	DO	mg/L	0	9.4	3.6	3	1.2	83.33
	BOD	mg/ L	6	48	11.8	9.53	1.6	80.76
	EC	µs/cm	84	980	436	239	1.12	54.8
	TS	mg/ L	100	900	359	144	0.3	40.11
	TDS	mg/ L	57	820	315	137	0.33	43.49
	Hardness	mg/ L	58.8	316	142.33	36.5	0.77	25.64
	pH	Unit	6.6	8.7	7.7	0.7	0.05	9.09
	No_3-N	mg/ L	0.01	14.02	1.9	2.26	0.71	118.7
	No_3-No_2	mg/l	0.03	8.07	0.89	1.64	1.32	179.7

The human activities and the nonlinear and complicated biochemical processes have a significant impact on the quality characteristics of water. Water quality variables such as dissolved oxygen (DO), biochemical oxygen demand (BOD), total solids (TS), alkalinity (Alk.), pH and electrical conductivity (EC) etc. all have an influence on the chemical oxygen demand (COD) of a given sample. For model development, it is vital to understand the relationship between these components and COD, like pH is a critical indicator of water quality because it is regulated by a series of interrelated chemical process that produce or utilize H^+ or OH^- ions. A low pH implies that organic compound is decomposing [39]. Various inorganic matter and salt ions dissolve in water and break-down into minute electrically charged particles called ions, hence boosting the

conductivity of water while lowering its solubility of oxygen. Similarly, when the salt content increases, the total dissolved solids in the river water stream increase, resulting in less atmospheric oxygen being dissolved. As noted, before, BOD and COD are indicators of the amount of biodegradable and non-biodegradable organic matter in a water sample that must be oxidized. As a consequence, total solids directly influence the COD content. Additionally, while nitrite is extremely toxic to aquatic life, it rapidly oxidizes to nitrate, which promotes the growth of water hyacinth, which covers the surface of a river, obstructing sunlight from reaching beneath the water's surface, reducing aeration and, as a result, increasing the demand for dissolved oxygen [39,40]. To select an adequate group of input variables from all possible variables is crucial for developing a high-quality model in any form of DDT model creation [4,5,11]. Numerous ways to choose input variables are explored in various studies. They are based on heuristics, expert knowledge, statistical analysis, or a combination of these [41–43]. However, there is a compelling need to do a thorough analysis of the input variable selection process. At the moment, there is no consensus on how this task should be carried out [44,45]. Thus, to identify input variables for the prediction of COD; expert knowledge, correlation factor (Table 1) and statistical analysis (Table 2) were utilized. A statistical approach has traditionally been used for providing a representative and reliable analysis of the water quality data. The statistical analysis of water quality characteristics reveals that the lowest values of all parameters were within the acceptable range; however, the maximum values were fairly high. In addition, the pattern of variation for the stretch between the Mula River and Mutha River was almost same, although substantial variations were detected in the Mula-Mutha River. Considering the above variation, correlation, and analysis, nine water quality parameters were identified for the prediction of COD for all three stretches. They are: BOD, DO, pH, hardness, electrical conductivity, total solids, total dissolved solids, nitrite, and nitrate (No3-No2/No3-N). Table 1 depicts the relationship between input parameters and COD for all three lengths of the Mula-Mutha River. The key contributing variables for COD in the Mutha and Mula rivers are BOD, DO, electrical conductivity (EC), total dissolved solids (TDS), total solids (TS), hardness (Hardn.) and nitrite (No3-N). However, for the Mula-Mutha River, a distinct pattern is found, with BOD, pH, nitrate (NO_3-NO_2), and DO as the key contributors following total solids, total dissolved solids and hardness, since nitrite quickly oxidizes to nitrate in the presence of dissolved oxygen, reflecting the breakdown or decomposition of organic matter (i.e., BOD).

5. Model development

The current study aims to explore the feasibility and effectiveness of DDT- MGGP and M5T to predict COD content for rivers: Mula, Mutha and Mula-Mutha. As discussed in the preceding section and illustrated in tables 1 and 2, all three stretches of river demonstrated distinct patterns of influence with respect to the input parameters because of different point and non-point sources of pollution and variation in the population nearby the river stretch. Hence, for Mutha river and Mula river commonly seven parameters were finalized out of nine i.e., BOD, DO, EC, hardness, TS, TDS, No3-N, while for Mula-Mutha River-BOD, DO, pH, hardness, TS, TDS, No3-No2

were finalized. Table 3 shows the model information for each set, including model numbers and input variables for each length of the Mula, Mutha, and combined Mula-Mutha River.

To predict COD, ANN's 3-layered FFBP model was trained to a low targeted error (mean squared error). ANN models were developed by utilizing the levernberg-marquardt algorithm to train the network using the "log-sigmoid" and "purelin" transfer functions. The data was standardized from 0 to 1. There are 7 neurons (or nodes) in the input layer, 1-15 neurons (or nodes) in the hidden layer, and 1 neuron (or node) in the output layer. The trial-and-error technique were utilized to determine the hidden nodes/neurons because finding a suitable number of nodes in the hidden layer is critical, as a greater number may result in over-fitting, meanwhile, a lesser number may not capture the information sufficiently [45]. All models were trained until a minimum-error target was reached, and their weights and biases were saved for testing on the remaining data sets. The ANN was trained in MATLAB 9.1 (2016) [45–47].

MGGP models were developed in the MATLAB 2016 with the source code of GPTIPS. The control parameters, such as the initial population, mutation frequency, and crossover frequency, were changed in every run to check the accuracy of the model. Every run was kept going until there was no more significant drop-in fitness, which meant generations without any more progress. In a single run, evolved programs are chosen for each data space at random. The parameters for a GP run were as follows: the population size was 500, the crossover rate was 84% and the mutation rate was 14%, The mean squared error was selected as the fitness criteria [25]. The best MGGP models were chosen based on high accuracy and less complexity. Pareto charts are also used to judge the complexity of a model. These charts show how the model evolved in terms of both complexity and accuracy. The MGGP algorithm was run several times with different populations and generations until the best model was found.

The M5 algorithm for MT was made with WEKA 3.9 [28]. The M5T technique employs the decision tree induction process to construct the tree, and the splitting function tries minimize the amount of intra-subset variation present on each level/branch of the tree. The tree was built using the standard deviation reduction factor (SDR) by decreasing intra-subspace variability and intra-subspace variability with a divide and conquer function [48]. The data of instances was divided into 70% for calibration, 15% for validation and 15% for testing respectively. Actual data was compared to projected or modelled values during the course of this investigation. Statistical indicators were utilised to assess the effectiveness of the developed DDT model e.g.: Root mean squared error (RMSE), Mean absolute relative error (MARE), and Coefficient of correlation (R) with visual presentation of values on graphs and scatter plots between observed and model-predicted values. RMSE and MARE should be as low as possible, and R ought to be close to 1, which is deemed to be the most accurate model [34]. For a model to be used in real life, it needs to be shown in a way that is easy to understand. For example, ANN shows the model as a set of trained weights and biases, M5T as a set of linear equations, and MGGP as a simplified equation. The MGGP method is known for the fact that the parameters that have little or no effect are removed from the final equation. In this way, these techniques can be used on-site and the confidence in the results can be raised.

Table 3
Detailed information of the developed model.

S. No.	Model No.	Station	Input variables	No. Data	Output variable
1	**ANN 1 MGGP 1 M5T 1**	Mutha (14km stretch)	BOD, DO, EC, Hardness, TS, TDS, No-N_3/ No3-N	144	COD
2	**ANN 2 MGGP 2 M5T 2**	Mula (30 km stretch)	BOD, DO, EC, Hardness, TS, TDS, No-N_3/ No3-N	206	COD
3	**ANN 3 MGGP 3 M5T 3**	Mula-Mutha (25 km stretch)	BOD, DO, pH, Hardness, TS, TDS, No-N_3/ No3-N	162	COD

5. Results and discussion

Table 4 displays the results of all three approaches (ANN, MGGP, and M5T) for predicting COD for the rivers Mutha, Mula, and Mula-Mutha. Individual set performances are addressed in depth further on. Table 5 illustrates the details of the best COD model in terms of the ANN model's architecture, the number of linear equations developed by M5T, and the parameters not included in the final developed equation by MGGP as output for the various models.

Table 4
RMSE, MARE & R for model developed using ANN, MGGP, and M5T for prediction of COD.

Targeted Output	Technique	Model No.	RMSE (mg/L)	MARE (mg/L)	R
COD	**ANN**	1 (Mutha)	0.078	0.006	0.92
		2 (Mula)	0.107	0.08	0.91
		3 (Mutha-Mula)	0.015	0.003	0.90
	MGGP	1 (Mutha)	0.68	0.009	0.91
		2 (Mula)	0.109	0.009	0.86
		3 (Mutha-Mula)	0.46	0.01	0.88
	M5T	1 (Mutha)	1.35	0.09	0.91
		2 (Mula)	2.23	0.19	0.90
		3 (Mutha-Mula)	1.98	0.09	0.89

Table 5
Detailed information of model developed using ANN-MGGP-MT.

Targeted output	Model No.	ANN-Architecture	Number of the equation in M5T	MGGP Parameters not considered in the equation
COD	1 (Mutha)	7:5:1	1	Hardness, TS, NO_3-NO_2
	2 (Mula)	7:8:1	3	TS, TDS, Hardness, NO_3-NO_2
	3 (Mutha-Mula)	7:1:1	1	TDS, Hardness, EC

The effectiveness of an ANN model is contingent upon how well the developed or constructed model is trained, and whether the synaptic weights and biases are appropriately adjusted to provide desirable output. This research makes use of the root-mean-squared error, or RMSE, since it illustrates the dispersion of the residual error (between observed and expected values), also known as the standard deviation of the residuals. The ANN model-3 (Mula-Mutha)

exhibited reasonable performance, with an RMSE of 0.015 mg/L and an R of 0.91. Similarly, ANN model-1 (Mutha) resulted in good performance with a RMSE of 0.078 mg/L and an R of 0.92. However, Model-2 (Mula) displays a high RMSE of 0.107 mg/L and R of 0.91. The higher RMSE is because the data are more dispersed, as seen by the larger deviation.

The subsequent approach used is MGGP, which provide the outcome as a streamlined equation-based model. MGGP instantly generates a mathematical-equation into a symbolic form that then can be analyzed to see how various parameters influence the final output and in which trend; this is the technique's distinctive feature (Refer equation 1).

$$COD = 8.72tanh(2.38NO_3 - NO_2^2TS^2) - 13.9tanh(tanh(2.45DONO_3 - NO_2)) + 11.2BOD^{1/2} + 5.81DO^{1/4} - 10.7 \quad (1)$$

In MGGP each gene or tree has a unique weighted coefficient [33]. Table 6 and figure 6 reflect that the statistical significance of genes 2, 4, and bias terms was greater than that of any other gene present. This means the input parameters included in genes 2 and 4 (DO, Nitrates, and Total solids) show a higher contribution or more influential toward the prediction of COD. To make the model simpler to use, the final outcome is presented in the form of an equation that is a linear sum of both the outputs and a bias value, the relative weights of which are indicated in Table 6 and illustrated in equation-1. The weight associated with each tree is derived using the least squares approach by trying to reduce/minimize the goodness of fit error between the predicted and the observed data. It is evident from equation 1 that the coefficient of weight for BOD, nitrates, and total solids is greater than that of DO. This result is consistent with the basic understanding of water quality acquired via environmental studies [49,50].

Table 6
Individual gene/tree weights (MGGP Model 3).

Term	Value
Bias	-10.7
Gene 1	5.81 $(BOD)^{1/4}$
Gene 2	-13.9 Tanh(Tanh(2.45(DO×NO_3-NO_2)))
Gene 3	8.72 Tanh(2.38(NO_3-NO_2)^2(TS)^2)
Gene 4	-11.2 $(BOD)^{1/2}$

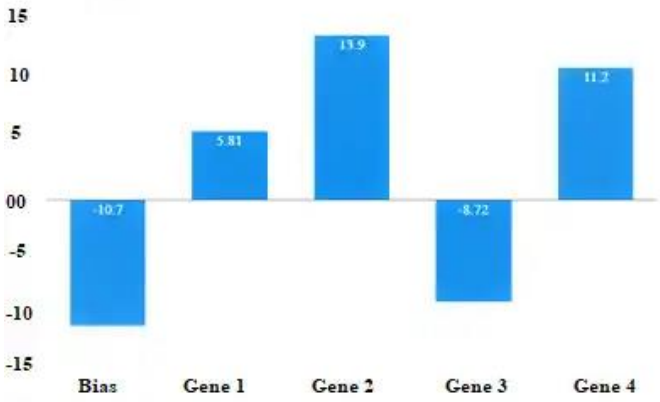


Fig. 6. Weights of the genes and bias (MGGP model 3).

Figure 7 and table 7 illustrates input frequency analysis, that is used to determine whether input variables are significant to the result for a particular model or a user-specified proportion of the whole population [47,51]. Three of the seven input variables in the MGGP model are significant: BOD, DO, and TDS, followed by hardness, TS and conductivity. MGGP is distinguished by its ability to reject input variables that do not significantly contribute to the final outcome.

Table 7
Individual Gene/Tree weights (MGGP Model 1).

Term	Value
Bias	-7.9
Gene 1	$-5.27\ e^{-6}$ (BOD) $(TDS)^2$
Gene 2	0.0471 $(BOD\text{-}DO)^2$
Gene 3	12.3 $(BOD)^{0.5}$
Gene 4	0.0058 (EC)-0.0058 (Hard.-7.45) (BOD-DO)

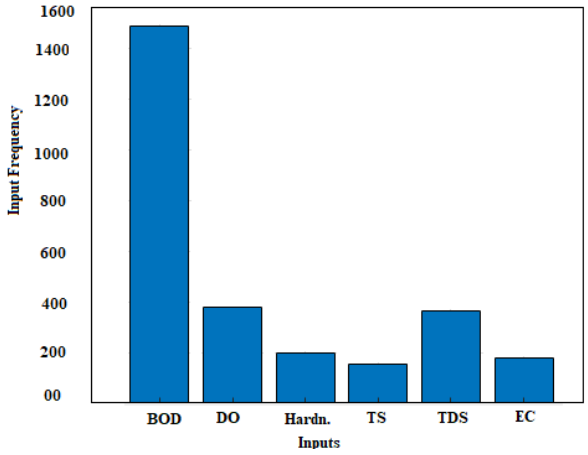


Fig. 7. Input frequency for MGGP model 1.

Table 1 also supports this statement as BOD, DO and TDS displayed the highest influence towards COD as compare to other input variables. This finding is compatible with recognized water quality concepts in environmental research [50]. Thus, one may conclude that MGGP's data-driven method has a decent grasp of the underlying COD phenomena and the correlation between the input and output parameters. In contrast to other methods, MGGP does not require a transfer function in order to develop successive generations of "offspring" according to the "specific fitness criteria" and genetic operations, allowing it to better detect and explore underlying patterns [5]. Additionally, the MGGP approach gives the developer a range of model possibilities. With the use of a Pareto chart, the best single model may be picked depending on the application requirements. The Pareto chart (Fig. 8) plots expressional complexity versus goodness of fit (R^2) or accuracy for models that are not dominated in terms of both complexity and performance by other solutions. A tree's complexity may be quantified by the number of nodes it contains, or by its expressive complexity [26,33]. Using a Pareto curve, a user may see the performance of solutions and pick a solution that maintains a balance between complexity and accuracy and its mathematical expression were presented in simplified equation [52]. The optimal model (high accuracy and simplicity) is denoted by a red dot/circle. Pareto (Green dots) models are those that are not substantially dominated by other models in terms of survival or fitness and complexity, while Non-Pareto (Blue dots) models are those that are strongly dominated.

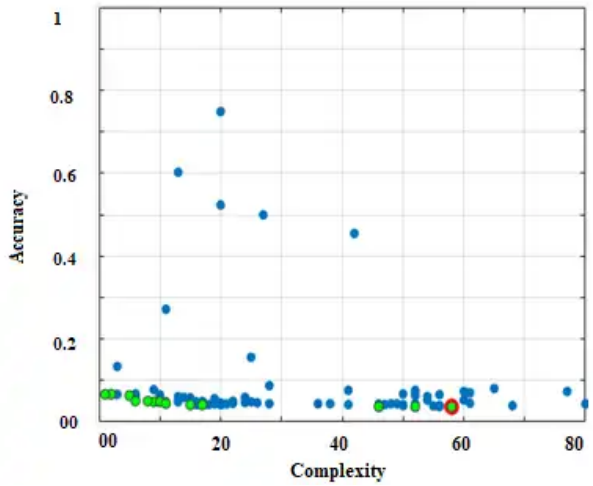


Fig. 8. Pareto front model report for MGGP model 1.

Model Tree is the third strategy used to estimate COD, and it is based on the divide and conquers principle. This technique measures the forecast value and sieves it on the routine route, smoothing it at each node in accordance with the linear node value anticipated for that node using the Linear Model. Figure 9 illustrates an example Model Tree generated using the M5 algorithm for M5T Model 2, along with developed linear regression equations. The below tree in figure 9 illustrates linear models (LM1-3) at various leaf nodes. The very first number inside the bracket indicates the number of related samples in the sorted subset of node, while the second number indicates the root mean square error (RMSE) of the associated linear model divided by the standard deviation of the sample's subset given in percentage [26]. Similarly, the number of equations derived for different models is shown in Table 5. Linear equations developed for M5T Model 2 using the M5 algorithm depict a negative coefficient for DO and a positive coefficient for all other parameters, particularly BOD and hardness, indicating that increasing the concentration of any of these impurities in a river increases the COD content, which is consistent with the theoretical understanding of their influence on COD [49]. Thus, it can be observed that M5T acquires a reasonable amount of knowledge about the underlying phenomena.

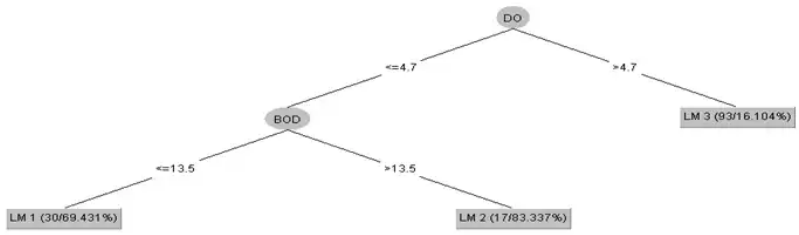


Fig. 9. Classifier tree for M5T model 2.

Linear equation developed for M5T Model 2 is as follows:

$$COD = 0.69BOD - 0.0765TDS + 0.0475EC + 0.0865\,Hardn. + 13.865 \quad (2)$$

$$COD = 1.40BOD - 5.755DO - 0.983TDS + 0.0596EC + 0.0565Hardn. + 9.7162 \quad (3)$$

$$CCOD = 1.99BOD - 1.033DO - 0.013TDS + 0.0067EC + 0.215Hardn. + 0.8904 \quad (4)$$

It is clearly visible from the scatter plot (fig.10) that the predicted COD values for model 1 was consistent with the actual values for all three developed models. All three data-driven strategies seem to have caught the fundamental phenomena. In scatter plot, the regression line supported by high value of correlation coefficient R (0.92, 0.91 and 0.91) respectively for ANN, MGGP and M5T. Further comparison between data-driven strategies using RMSE reflects that ANN have improved the performance of the proposed model as it has lower value of RMSE (0.078). The RMSE is a statistic that measures the average difference between the model's predicted values and the data set's actual values. The amount to which RMSE surpasses is a measure of the presence of outliers in the data. The ANN algorithm creates an approximation function that matches chosen input parameters to the intended outputs, which is then evaluated. As part of the process, it makes adjustments to weights and biases to achieve the expected goal, which results in a more flexible approach. While considering Eq. 5 for M5T model 1, showing direct contribution of only three main input parameters i.e., BOD, TDS, and TS with COD, which is also supported by fundamentals of water quality as COD, reflects the total organic matter in water that is contributed by biodegradable matter as well as total solids too [49]. Hence M5T Model 1 shows good performance as compare to M5T Model 2 &3 in terms of RMSE.

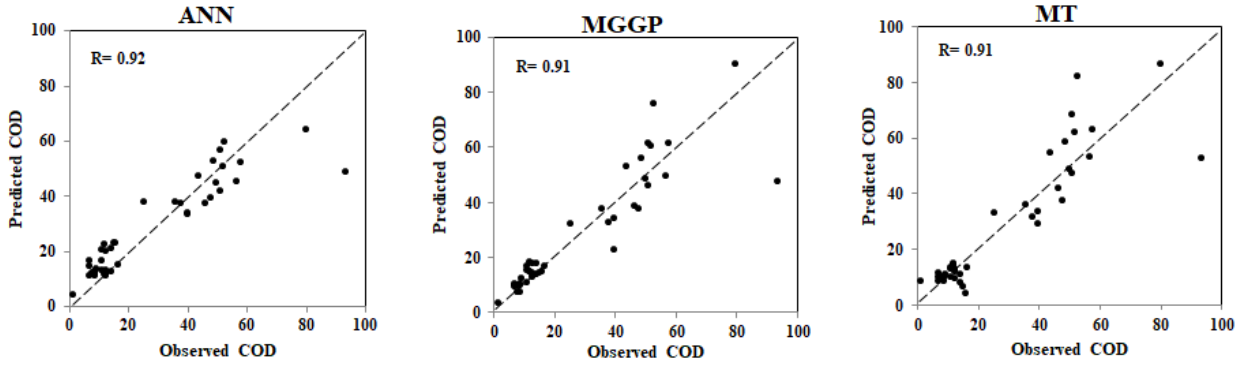


Fig. 10. COD model-1 by ANN-MGGP-M5T.

$$COD = 2.234BOD + 0.0432TS + 0.0726TDS + 7.7016 \quad (5)$$

The scatter plot (figure 11) shows that ANN, MGGP, and M5T for model 2, all had a balanced dispersion, except for a few high COD values recorded by the MGGP model with the lowest R = 0.86. In addition, the RMSE value for ANN (0.107) is the lowest, followed by MGGP (0.109) and M5T (2.23). Eq. 6 generated by MGGP for model 2, emphasizes that it has considered only three parameters (out of seven), namely BOD, DO, and EC, while rejecting all other parameters; this may be the explanation for the model's poor performance, since TDS, TS, and hardness contribute to an increase in organic matter (non-biodegradable) in water and also show a good correlation with COD. Similarly, model trees don't employ all of the parameters included as input variables to create linear models (LM) at each leaf node. Only those parameters that meet the constraints of particular criteria (standard deviation reduction) fall under one sub-tree, which

ends in a leaf node, which is also reflected in eq. 2-4. Thus, this quality of M5T makes the model outperform in terms of R (0.90).

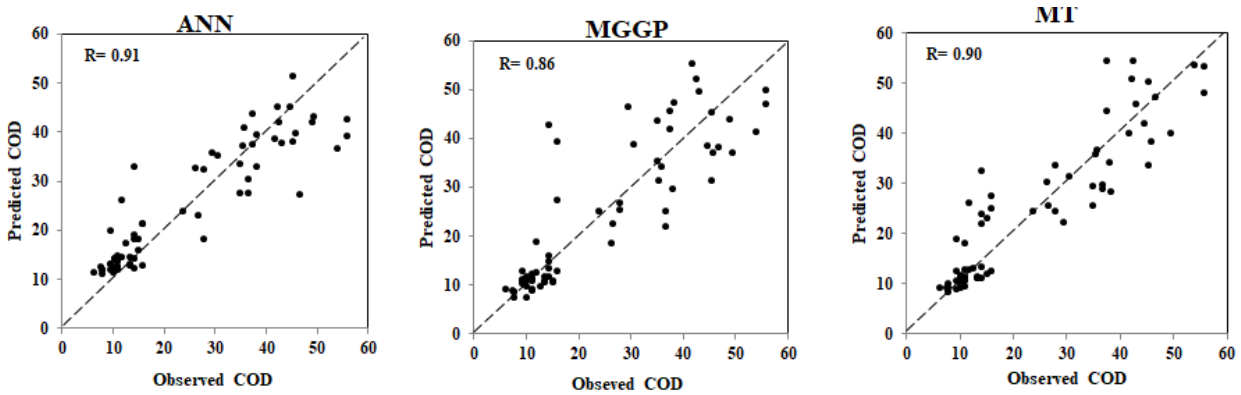


Fig. 11. COD model-2 by ANN-MGGP-M5T.

$$1.26e^{-4}\,(EC)^2\,tanh\,(DO) + 3.9\,(BOD)^{3/4}\,tanh(BOD) - 3.13e^{-5}\,(DO)\,(EC)^2 + 8.29e^{-4}\,(DO)^2\,(EC) + 1.39 \quad (6)$$

ANN, MGGP, and M5T all showed good performance in terms of R as shown in the scatter plot (figure 12) for model 3 (0.90-0.89). In terms of RMSE, however, the ANN and the MGGP model came out on top (0.015 - 0.46). Table 2 demonstrate that the standard deviation of Mula and Mutha data is lower than that of the combined Mula-Mutha dataset (model 3). Thus overall ANN technique displays similar values as observed values which can be seen through lower RMSE (0.078 with ANN, 0.680 with MGGP and 1.35 with MT) with Mutha river and other rivers as well (refer table 4). All these models with ANN, MGGP and MT also support their performance with lower MARE values (0.006 with ANN, 0.009 and 0.09 with MT) and higher r value (0.92 with ANN, 0.91 by MGGP and 0.91 with M5T). The similar trend is seen in models developed for Mula and Mula-Mutha river. Lower RMSE shows weighted measure of the error in which the standard deviation contributes the most between observed and modelled values. RMSE is sensitive towards higher values, MARE shows its inclination towards lower predicted values and r value towards the higher values. The higher value observed for Mula river is 56 and predicted by ANN is 56.433, 64.545 by MGGP and 65.832 by MT. This trend is also seen in other developed models as well. Thus, ANN predicts the values closer to observed values followed by model developed using MGGP and then by M5T. P value analysis is another statical method used to accept or reject the null hypothesis which states that there is no anomaly between the observed value and predicted value of ANN, MGGP and M5T respectively. The p value for Mula river is calculated as 0.551 with ANN, 0.981 with MGGP and 0.342 with M5T. P values for Mutha river with ANN:0.763, MGGP:0.997 and M5T:0.870 along with p values for Mula-Mutha river (0.810 with ANN, 0.960 with MGGP and 0.741 with M5T). These p values for Mula, Mutha and Mula-Mutha river are greater than 0.05 thus indicating the failure to reject the null hypothesis.

It is also dicussed in above section and reflected in table 1, that major contributing parameters for Mula river and Mutha river where DO, BOD, EC, TS and TDS while for Mula-Mutha Stretch, BOD, pH, nitrate (NO_3-NO_2), and DO were found to be major contributing parameters. These discrepancies in data with its variation might be explained by variations in regional

climate and pollution patterns. The Divide and conquer approach used by M5T assist the leaf node to determine splitting criteria by minimising the standard deviation. As already discussed in the preceding section, the standard deviation of the 3rd stretch, i.e., Mula-Mutha, is on the higher side, which indicates more spread of values and contributes to the larger margin error. Therefore, M5T displays RMSE as 1.98, which is on the higher side than expected. Overall, ANN and MGGP performed better than M5T for the majority of segments. When the current study's results are compared to previous findings, it is noticed that Murat [53] utilized ANFIS to forecast COD for a waste water treatment plant and obtained a minimum RMSE of 54.9 mg/L, which is rather high when compared to current results; in contrast, Olyaie et al. [40] and Heddam [9] used ANN, LSVM, and LGP to predict the water quality parameter DO and obtained RMSEs of 0.592 mg/L, 0.645 mg/L, and 0.374 mg/L, respectively, exhibiting a similar pattern as shown here.

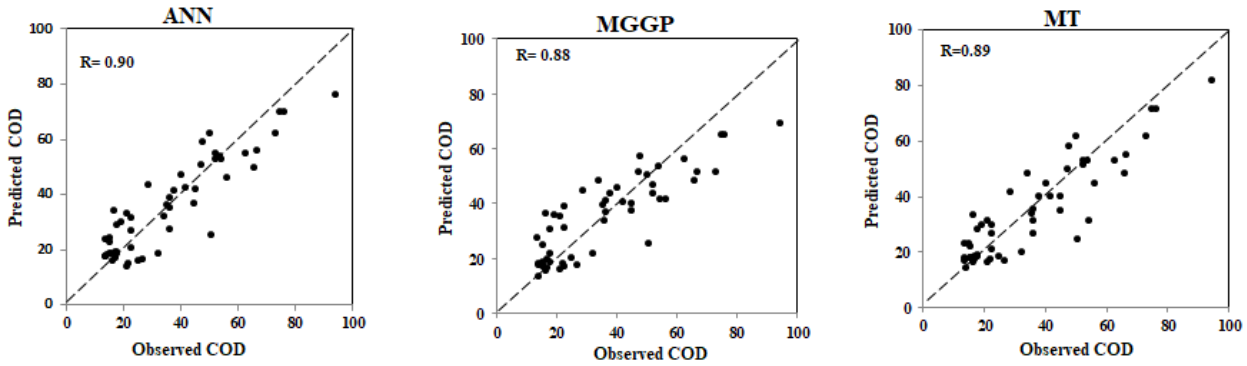


Fig. 12. COD model-3 by ANN-MGGP-M5T.

7. Conclusion

In recent years, it has been shown that data-driven strategies perform best in complex, non-linear, and unexpected environments. Analyzing environmental data using data-driven approaches like ANN, MGGP, and M5T is the focus of this research. Using DDT in environmental research has progressed over the last two decades, but the results also reveal areas that need more investigation to improve water quality management. The present study highlighted the potential of DDT in predicting water quality measures that are difficult to measure in the field. It's reasonable to say that all of the models for the prediction of important water quality parameter COD for all the three stretches of the river Mula-Mutha did well enough. ANN-developed models outperformed MGGP and M5T with higher R and lower RMSE values. In terms of accuracy, resilience, and fault tolerance, ANN excels all other analytical approaches because of its model-free structure and capacity to map nonlinear input-output relationships. The MGGP models 2 and 3 display better performance by providing RMSE values of less than 0.5 mg/L (0.109, 0.46), and R values of more than 0.85 (0.86, 0.88). When compared to the M5T models, the MGGP models exhibited a considerable reduction in their RMSE values, which is an indication of significant progress. It would seem that the outcomes of all three models are impacted by the data's inherent variability. Findings from this research also show that these three models can learn from examples and display input parameters that are in sync with the environmental studies domain knowledge that they were designed to learn from. It is also

suggested that the models be developed by combining data from all three stretches into a single dataset, since this would give more data to train and perhaps increase accuracy.

Acknowledgments

I express my sincere gratitude to Nashik Hydro Work, Maharashtra for providing us water quality data for my research work. I am thankful to Dr. Pradnya Dixit Associate Professor VIIT college for the help and moral support to carry out the research.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

Authors contribution statement

P.S. (Ph.D. Scholar) Conducted all the data collection, model formation, calculations and wrote the manuscript. S.N.L (Professor) revised the manuscript. P.S.K. (Associate Professor) revised the manuscript.

References

[1] Zare Abyaneh H. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J Environ Heal Sci Eng 2014;12:40. https://doi.org/10.1186/2052-336X-12-40.

[2] Verma AK, Singh TN. Prediction of water quality from simple field parameters. Environ Earth Sci 2013;69:821–9. https://doi.org/10.1007/s12665-012-1967-6.

[3] Jingsheng C, Tao Y, Ongley E. Influence of High Levels of Total Suspended Solids on Measurement of Cod and Bod in the Yellow River, China. Environ Monit Assess 2006;116:321–34. https://doi.org/10.1007/s10661-006-7374-2.

[4] Rice EW, Bridgewater L, Association APH. Standard methods for the examination of water and wastewater. vol. 10. American public health association Washington, DC; 2012.

[5] Londhe SN, Panchang V. Correlation of wave data from buoy networks. Estuar Coast Shelf Sci 2007;74:481–92. https://doi.org/10.1016/j.ecss.2007.05.003.

[6] Palani S, Liong S-Y, Tkalich P. An ANN application for water quality forecasting. Mar Pollut Bull 2008;56:1586–97. https://doi.org/10.1016/j.marpolbul.2008.05.021.

[7] Singh KP, Basant A, Malik A, Jain G. Artificial neural network modeling of the river water quality—A case study. Ecol Modell 2009;220:888–95. https://doi.org/10.1016/j.ecolmodel.2009.01.004.

[8] Akilandeswari S, Kavitha B. Comparison of ANFIS and statistical modeling for estimation of chemical oxygen demand parameter in textile effluent. Der Chem Sin 2013;4:96–9.

[9] Heddam S, Kisi O. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 2018;559:499–509. https://doi.org/10.1016/j.jhydrol.2018.02.061.

[10] Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environ Model Softw 2000;15:101–24. https://doi.org/10.1016/S1364-8152(99)00007-9.

[11] Danandeh Mehr A, Ghadimi S, Marttila H, Torabi Haghighi A. A new evolutionary time series model for streamflow forecasting in boreal lake-river systems. Theor Appl Climatol 2022;148:255–68. https://doi.org/10.1007/s00704-022-03939-3.

[12] Karami H, Ghazvinian H, Dehghanipour M, Ferdosian M. Investigating the Performance of Neural Network Based Group Method of Data Handling to Pan's Daily Evaporation Estimation (Case Study: Garmsar City). J Soft Comput Civ Eng 2021;5:1–18. https://doi.org/10.22115/scce.2021.274484.1282.

[13] Najah A, El-Shafie A, Karim OA, El-Shafie AH. Application of artificial neural networks for water quality prediction. Neural Comput Appl 2013;22:187–201. https://doi.org/10.1007/s00521-012-0940-3.

[14] Basant N, Gupta S, Malik A, Singh KP. Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water — A case study. Chemom Intell Lab Syst 2010;104:172–80. https://doi.org/10.1016/j.chemolab.2010.08.005.

[15] Elmolla ES, Chaudhuri M, Eltoukhy MM. The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. J Hazard Mater 2010;179:127–34. https://doi.org/10.1016/j.jhazmat.2010.02.068.

[16] Emamgholizadeh S, Kashi H, Marofpoor I, Zalaghi E. Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. Int J Environ Sci Technol 2014;11:645–56. https://doi.org/10.1007/s13762-013-0378-x.

[17] Wu X, Zhang Q, Wen F, Qi Y. A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China. Water 2022;14:3408. https://doi.org/10.3390/w14213408.

[18] Ozkan O, Ozdemır O, Azgın ST. Prediction of biochemical oxygen demand in a wastewater treatment plant by artificial neural networks. Asian J Chem 2009;21:4821–30.

[19] Dogan E, Sengorur B, Koklu R. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. J Environ Manage 2009;90:1229–35. https://doi.org/10.1016/j.jenvman.2008.06.004.

[20] Danandeh Mehr A, Safari MJS. Multiple genetic programming: a new approach to improve genetic-based month ahead rainfall forecasts. Environ Monit Assess 2019;192:25. https://doi.org/10.1007/s10661-019-7991-1.

[21] Ay M, Kisi O. Modeling of Dissolved Oxygen Concentration Using Different Neural Network Techniques in Foundation Creek, El Paso County, Colorado. J Environ Eng 2012;138:654–62. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000511.

[22] ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. Artificial Neural Networks in Hydrology. I: Preliminary Concepts. J Hydrol Eng 2000;5:115–23. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(115).

[23] Dawson CW, Wilby RL. Hydrological modelling using artificial neural networks. Prog Phys Geogr Earth Environ 2001;25:80–108. https://doi.org/10.1177/030913330102500104.

[24] Jain P, Deo MC. Neural networks in ocean engineering. Ships Offshore Struct 2006;1:25–35. https://doi.org/10.1533/saos.2004.0005.

[25] Shahin MA. State-of-the-art review of some artificial intelligence applications in pile foundations. Geosci Front 2016;7:33–44. https://doi.org/10.1016/j.gsf.2014.10.002.

[26] Quinlan JR. Learning with continuous classes. 5th Aust. Jt. Conf. Artif. Intell., vol. 92, World Scientific; 1992, p. 343–8.

[27] Kulkarni P, Londhe SN, Dixit PR. A comparative study of concrete strength prediction using artificial neural network, multigene programming and model tree. Chall J Struct Mech 2019;5:42. https://doi.org/10.20528/cjsmec.2019.02.002.

[28] Hashmi S, Halawani SM, Barukab OM, Ahmad A. Model trees and sequential minimal optimization based support vector machine models for estimating minimum surface roughness value. Appl Math Model 2015;39:1119–36. https://doi.org/10.1016/j.apm.2014.07.026.

[29] Abolfathi S, Yeganeh-Bakhtiary A, Hamze-Ziabari SM, Borzooei S. Wave runup prediction using M5′ model tree algorithm. Ocean Eng 2016;112:76–81. https://doi.org/10.1016/j.oceaneng.2015.12.016.

[30] Solomatine DP, Xue Y. M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. J Hydrol Eng 2004;9:491–501. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491).

[31] Solomatine DP, Dulal KN. Model trees as an alternative to neural networks in rainfall-runoff modelling. Hydrol Sci J 2003;48:399–411. https://doi.org/10.1623/hysj.48.3.399.45291.

[32] Searson DP, Leahy DE, Willis MJ. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. Proc. Int. multiconference Eng. Comput. Sci., vol. 1, Citeseer; 2010, p. 77–80.

[33] N. S, R. P. Genetic Programming: A Novel Computing Approach in Modeling Water Flows. Genet. Program. - New Approaches Success. Appl., IntechOpen Publishing London, UK; 2012. https://doi.org/10.5772/48179.

[34] Gandomi AH, Alavi AH. A new multi-gene genetic programming approach to nonlinear system modeling. Part I: materials and structural engineering problems. Neural Comput Appl 2012;21:171–87. https://doi.org/10.1007/s00521-011-0734-z.

[35] Pune Municipal Corporation. JICA Project | Pune Municipal Corporation 2023. https://www.pmc.gov.in/en/jica-project.

[36] Report On Environmental Status of Pune Region. Kalpataru Point, Sion Circle, Sion (East) Mumbai. n.d.

[37] Sahu P, Karad S, Chavan S, Khandelwal S. Physicochemical Analysis Of Mula Mutha River Pune. Civ Eng Urban Plan An Int J 2015;2.

[38] Central Pollution Control Board (CPCB) | The Official Website of Ministry of Environment, Forest and Climate Change, Government of India n.d.

[39] Fletcher D, Goss E. Forecasting with neural networks. Inf Manag 1993;24:159–67. https://doi.org/10.1016/0378-7206(93)90064-Z.

[40] Olyaie E, Zare Abyaneh H, Danandeh Mehr A. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. Geosci Front 2017;8:517–27. https://doi.org/10.1016/j.gsf.2016.04.007.

[41] Ranković V, Radulović J, Radojević I, Ostojić A, Čomić L. Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. Ecol Modell 2010;221:1239–44. https://doi.org/10.1016/j.ecolmodel.2009.12.023.

[42] Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Heal 2020;8. https://doi.org/10.1136/fmch-2019-000262.

[43] Adeniran KA, Adelodun B, Ogunshina M. Artificial Neural Network Modelling of Biochemical Oxygen Demand and Dissolved Oxygen of Rivers: Case Study of Asa River. Environ Res Eng Manag 2017;72:59–74. https://doi.org/10.5755/j01.erem.72.3.14120.

[44] G. E. McCuen. Protecting water quality 1986:180.

[45] Hem JD. Study and interpretation of the chemical characteristics of natural water. US Geological Survey; 1959. https://doi.org/10.3133/wsp1473_ed1.

[46] MathWorks Announces Release 2016b of the MATLAB and Simulink Product Families - MATLAB & Simulink. MathWorks 2016.

[47] Singh HK. Prediction of shear strength of deep beam using Genetic Programming 2014.

[48] Melesse AM, Khosravi K, Tiefenbacher JP, Heddam S, Kim S, Mosavi A, et al. River water salinity prediction using hybrid machine learning models. Water 2020;12:2951.

[49] Garg SK. Water Supply Engineering. Khanna Publishers; 2010.

[50] Metcalf E. Wastewater Engineering Treatment and Reuse (4th edition) (2004) | Akhid Maulana - Academia.edu. 4th editio. 2004.

[51] Rahimikhoob A, Behbahani SMR, Banihabib ME. Comparative study of statistical and artificial neural network's methodologies for deriving global solar radiation from NOAA satellite images. Int J Climatol 2013;33:480–6. https://doi.org/10.1002/joc.3441.

[52] Danandeh Mehr A, Jabarnejad M, Nourani V. Pareto-optimal MPSA-MGGP: A new gene-annealing model for monthly rainfall forecasting. J Hydrol 2019;571:406–15. https://doi.org/10.1016/j.jhydrol.2019.02.003.

[53] Ay M, Kisi O. Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. J Hydrol 2014;511:279–89. https://doi.org/10.1016/j.jhydrol.2014.01.054.