Water Quality Monitoring using Data Mining Techniques

Mitali Kathpal* and Kriti Saroha**
*-**School of Information and Technology
mitalikathpal@outlook.com, kritisaroha@cdac.in

Abstract: This paper conducts a study, where data mining techniques were applied to river quality monitoring database. The effects of water quality deterioration can be handled properly if water quality is predicted beforehand. In this, some techniques have been introduced to examine the future quality of water. A total of 8451 samples of water quality were collected with 1 hour time interval from different sites to examine the water quality. This paper includes the values of fifteen parameters which affect the quality of water. Also, the impact of weather on water quality is studied by examining the interdependence between water quality attributes and weather parameters.

Keywords: ANN with NAR, NARX, relief, regression, WQI.

Introduction

Water quality is becoming a major issue now-a-days as the water is getting polluted because of numerous factors that includes industrial as well as commercial movements and also human and natural actions, also poor sanitation infrastructure. The water contamination is adversely affecting health, environment and infrastructure. The focus behind the research is to introduce a method to analyze and predict the quality of water of selective regions with the help of water quality parameters. There are some biological, physical and chemical parameters that affect water quality. (approx. 15 in the dataset)

This paper addresses the problem by introducing a model based on Machine Learning Techniques for predicting the future water quality trends with help of historic water quality data of that region. Artificial Neural Network with nonlinear autoregressive model is used to make effective prediction of water quality trends.

Also, the approach is extended further by determining the correlation between water quality and weather attributes. This helps in analyzing the impact of weather on water quality. In the proposed approach, the ranking algorithm has been used to decide the importance of each predictor in determining the water quality. The remainder of the paper is organized in the following manner. In section 2, is discussed the related work in the area. Section 3 describes the proposed approach. Section 4 summarizes the conclusion and results.

Related Works

Various researches have been done to extract the pattern trends that can examine the water contamination.

Some of the techniques used by different researchers are briefly described in this section.

Ref.[1] Yafra khan and Chai Soo see proposed ANN with NAR time series model for water quality prediction. The data was collected from U.S Geological Survey's(USGS) National Water Information System(NWIS) with 6 minute time interval for efficient prediction of water quality. The performance of the model was depicted through Regression, Mean Squared Error (MSE) and Root Mean Squared Error(RMSE). The model predicted improved results with lowest MSE for turbidity 3.7*10^-4 and best regression value 0.98 for specific conductance.

Ref.[6] Ting nien Wu and Chiu Sheng Su proposed a method for knowledge discovery which is presented in Fig 1. The data was accumulated from Eighty-four monitoring wells located in Taiwan.

The data set has been distributed by clustering into 35% of the monitoring wells for salinization, 11% for arsenic dissolution,19% for organic pollution, and 35% for mineralisation.[2]

Ref. [2] Jan-Yee-Lee proposed a model known as GREY MODEL for prediction based on single variable. The dataset was collected from 84 monitoring wells in chinanian Blackfoot disease region from 2009 to 2012. The collected dataset included concentration of arsenic in groundwater which was taken as input for model. The method proposed by author have these steps given in Fig 2.16 wells out of 84 monitoring wells in chianian plain groundwater subregion, have high arsenic level as concluded by grey model. The prediction model based on grey theory have improved the accuracy of limited data prediction and forecasting. This model helped in improvement of groundwater monitoring.

Ref. [4] Ivana.D.Radojevic, et.al. used clustering using k-means and classification with decision trees for prediction of state of coliforms and water quality monitoring. This approach proposed by author proves to be reliable one in predicting the total coliforms in the dataset. A total of 172 samples were considered for analysis.

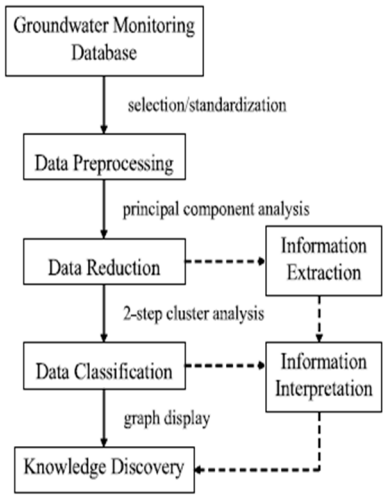

Fig 1

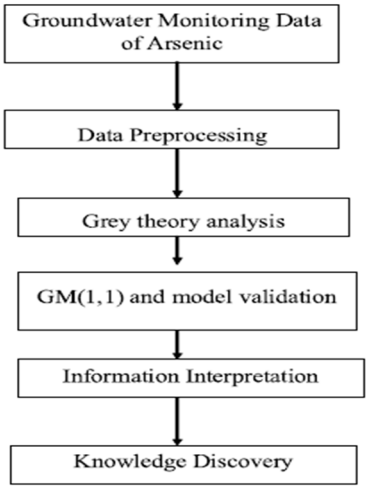

Fig 2

Ref. [3] Maqbool Ali and Ali Mustafa Qamar collected an entire set of 663 samples from distinct locations from 2009 to 2012. The author proposed methodology which consists of five parts. Wherein, the first part, presents month-wise quality parameter trends and the second part provides the parametric satisfactory analysis. In third part, the data is pre-processessed and outliers are removed. The fourth part computes the best quality index with the help of clustering methods and the last phase determines the months when there is high contamination of fecal coliforms. The results analysed in this paper were as follows:- it was concluded that Hierarchical clustering with Avg. Linkage method using Euclidean distance to calculate water quality index have given good results than any other techniques. And for classification, MLP has provides better output. It was found that the concentration of fecal coliforms was more in months of june,july and october.

Ref. [5] Neetu Arora, Amarpreet Singh Arora, Siddhartha Sharma and Dr. Akepati S. Reddy presented a approach which is useful in understanding the main pollutants in water quality deterioration. In this case study proposed by author, cluster analysis was used to study the temporal and spatial variation in surface water quality of satluj river. It helps in reducing the number of monitoring stations.

The comparison of various approaches described in this paper is given in table 1.

Proposed Approach

This section presents the proposed approach that would be used to extract knowledge from water quality dataset.

The steps of proposed approach shown in Fig 3 are explained below:-

The proposed approach is divided into three parts. In the first part the class labels are defined using water quality index calculation.

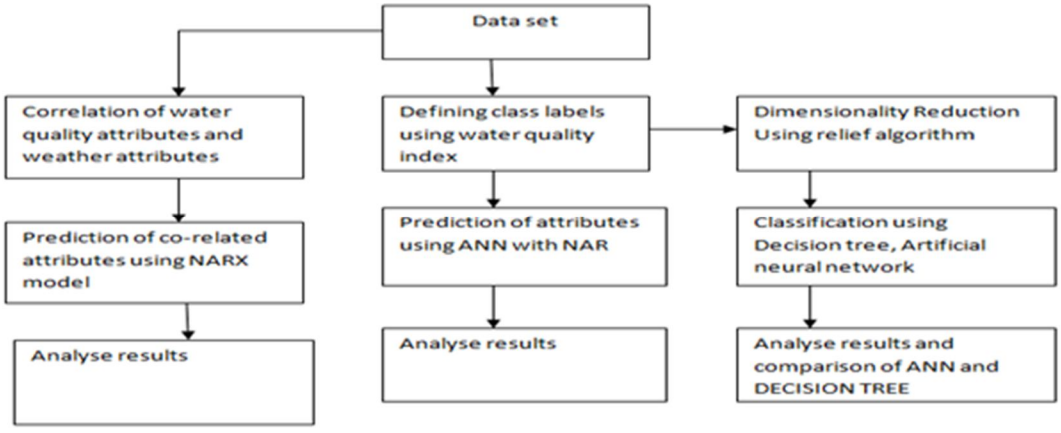

Figure 3: Proposed Approach

Table 1: Comparison Table

	Yafra khan; Chai Soo See	Ting nien wu, et.al.	Jan-Yee-Lee	Ivana.D.Radojevic, et.al.	Maqbool Ali, et.al.	Neetu Arora, et.al.
Algorithm Used	Artificial neural networks	PCA Cluster Analysis Using K-MEANS	Grey model	Clustering and classification using decision trees	Classification techniques MLP,RBF,KNN,SVM Clustering techniques k-means Hierarchial	Hierarchical clustering
Attributes	4 attributes	14 attributes	Only arsenic	Only coliforms	12 Attributes	8 monitoring stations
Conclusion	This model proves reliable with high prediction accuracy. High regression value for specific conductance being 0.98	The clustering divides the dataset as follows:- 35% for salinization, 11 % for arsenic dissolution, 19% for organic pollution, and 35% for mineralization	Accuracy level is good with error<0.01	Better prediction of coliforms in water quality to maintain ecological balance.	High accuracy with prediction that concentration of fecal coliforms were more during july, march, june and October.	Better understanding of parameters that are most important for water quality.
Future Work	User centric approach	-	-	-	Time Series forecasting model	Principal component analysis

WQI is computed to define the rating of quality whether it is good or bad based on
water quality parameters. Equation (iii) is used to compute the Water Quality Index. Equation (i) is applied to calculate the Unit weight (Wi).

$$Wi=k/Si \quad \text{.....(i)}$$

Where *k* = Proportionality constant

$$K= 1/(\sum(1/Si) \quad \text{.....(ii)}$$

Si = Standard permissible value of *ith* parameter given by WHO standard.
Wi = unit weight of ith parameter
Water quality index (Weighted Arithmetic Mean method)

$$WQI=\sum qiwi/\sum wi \quad \text{.....(iii)}$$

Where qi= Sub Index the quality rating for ith parameter

$$Qi=100*(Vi/Si) \quad \text{.....(iv)}$$

Vi = observed value in laboratory; n = total no. of parameters used; Si = standard value of ith parameter. Secondly, predicts the water quality trends of 15 water quality parameters using ANN with NAR model. The dataset used for analysis is with 1 hour time interval. At the final stage of this part of approach, the results are analysed. In the second part of the approach, ranking algorithm is applied to establish the importance of each attribute in determining the water quality. In the second ANN and DECISION TREE classifiers are applied. In the last step, the results of classifiers are analyzed and performance of both classifiers are compared.
The third part of the approach determines the correlation between weather and water attributes to study the interdependence between these two datasets. Also, prediction of co-related attributes have been found using ANN with NARX model. At last step, the results are analyzed.

Results & Conclusion

The performance of ANN with NAR time series model is measured through Regression value, mean squared error. The proposed model that includes ANN-NAR has proven to be efficient giving the prediction accuracy with lowest MSE being 0.0006 for phosphate and the best regression value for pH 0.87492.The graphs generated for regression analysis indicate how strongly the data fits the function for training validation and testing. The regression value closer to 1, indicates that the

function fits better and has good accuracy of prediction. The graphs for MSE shows the measure of epochs taken to converge testing, validation and training.

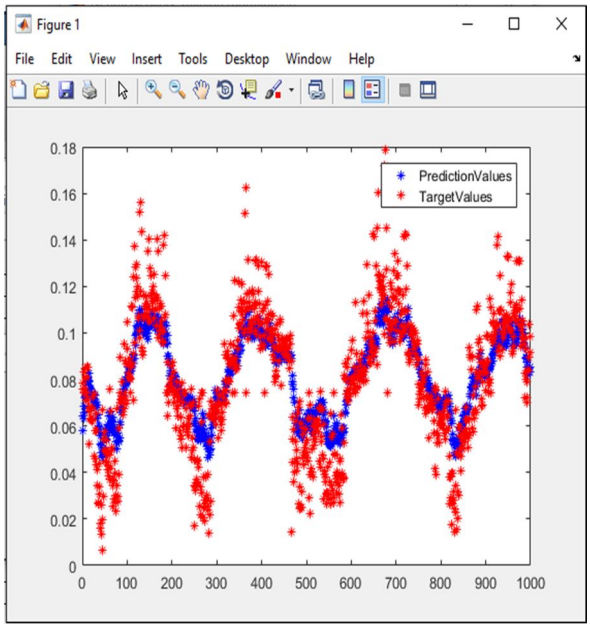

Fig 4: Prediction of Ph using ANN with NAR

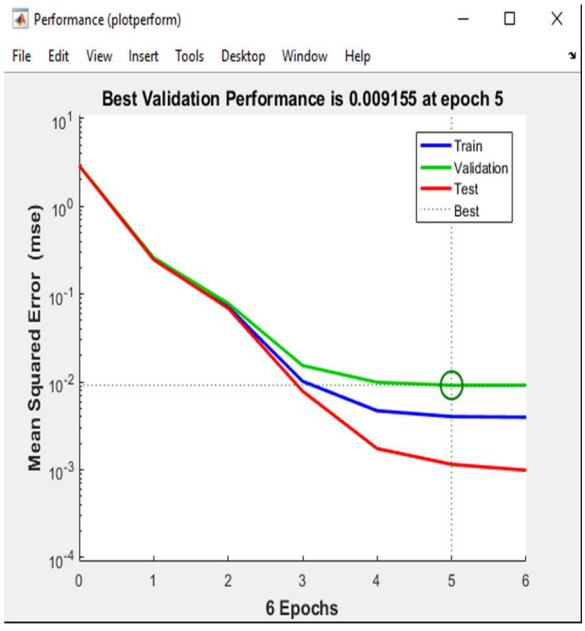

Fig 5: Mean Squared Error graph

The blue dots in the Fig 4 shows the target values and the red dots shows the predicted values by the proposed model.

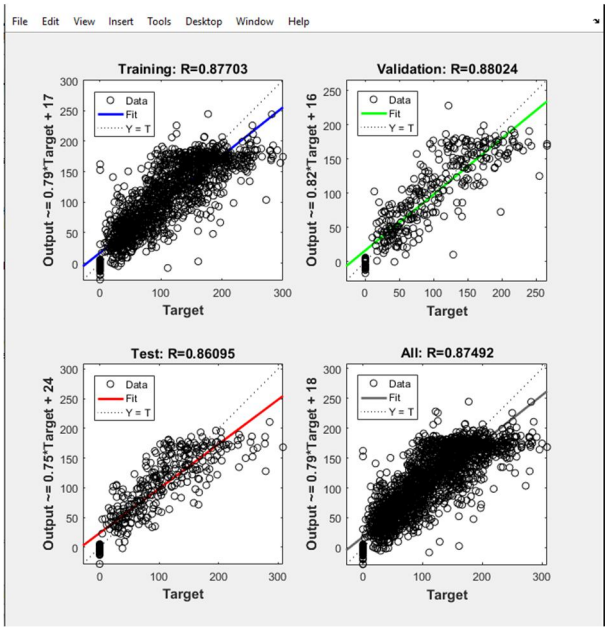

Fig 6: Regression Analysis

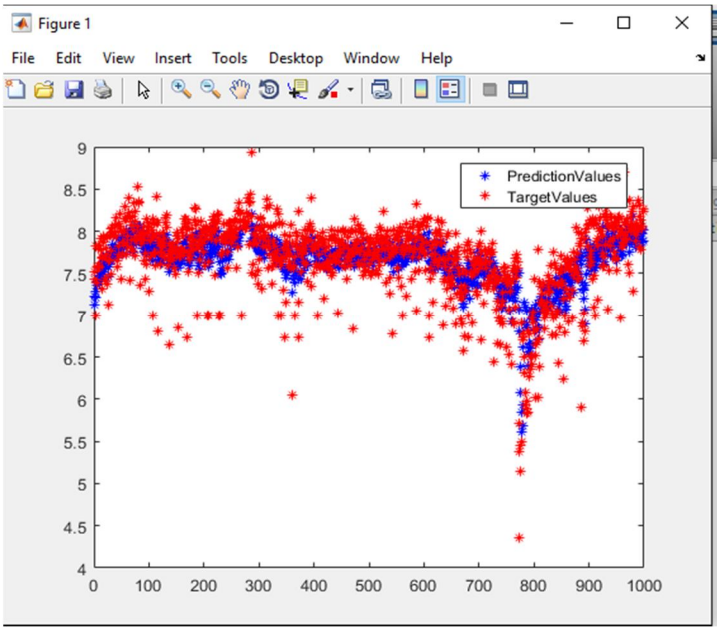

Fig 7: Prediction of PHOSPHATE using ANN with NAR

Fig 6 shows in what way the model fits the function for all validation, testing and training.

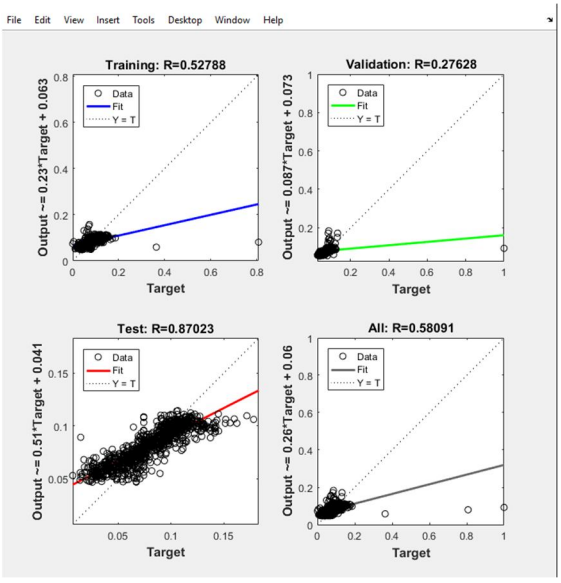

Fig 8: Regression Analysis

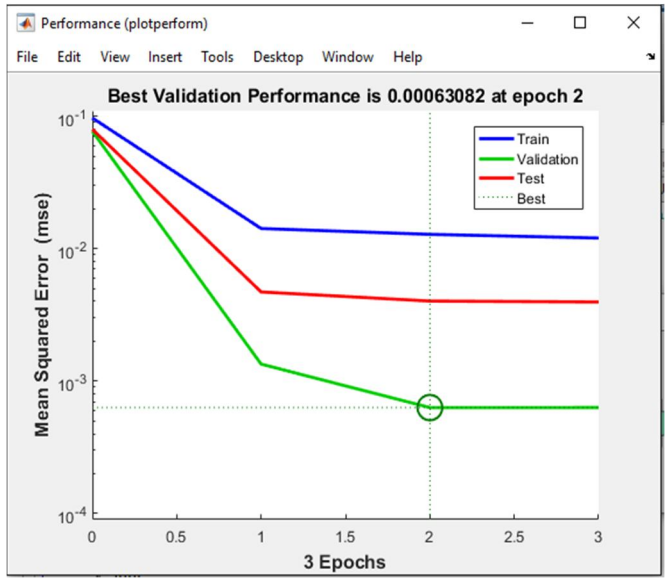

Fig 9: Mean Squared Error

Fig 9 represents the best validation performance is at epoch 5 where training, validation and testing line converge

Comparison of Results

Table 2: comparison of results using ANN with NAR model

ATTRIBUTE NAME	REGRESSION VALUE	MSE VALUE
PH	0.87492	0.009155
CONDUCTIVITY	0.58	0.007658
DISSOLVED OXYGEN	0.58091	2.6906e-08
TURBIDITY	0.6263	0.021035
NITRITE	0.6345	1.5695e-05
NITRATE	0.30305	0.008596
TOTAL COLIFORMS	0.31438	1.3511e-05
E-COLI	0.6189	3.1997e-05
FAECAL STREPTOCOCCI	0.31438	3.1997e-05
B.O.D	0.021035	0.021035
C.O.D	0.68581	3.1997e-05
PHOSPHATE	0.5809	0.00063082
SALINITY	0.7415	0.0079619
AMMONIUM	0.6513	0.005879

In the second part of the approach, the ranking of the attributes is determined by relief algorithm. This algorithm gives the ranking to each attribute by examining how strongly it determines the quality of water. The ranking is shown in Fig 10.
The figure shows that attribute 13 has highest ranking and attribute 6 has lowest ranking.

Variables - ranked

ranked

1x15 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	13	8	9	4	3	15	1	2	12	14	10	5	7	11	6
2															

Fig 10: Ranking of attribute

Classification using ANN results are determined by regression analysis plot. The first plot shows the regression analysis with all the 15 attributes. The value of regression analysis is 0.90967 which is shown in Fig 11.
The second regression analysis plot shows the plot after removal of nitrite attribute that is the lowest ranking attribute. It is analyzed that the value of regression increases from 0.90967 to 0.91506 shown in Fig 12.
The third regression analysis plot shows the plot after removal of ammonium attribute that is the second lowest ranking attribute. It is analyzed that the value of regression increases from 0.91506 to 0.93743 shown in Fig 13.
The fourth regression analysis plot shows the plot after removal of turbidity attribute i.e the third lowest ranking attribute. It is analyzed that the value of regression increases from 0.93743 to 0.90078 shown in Fig 14.

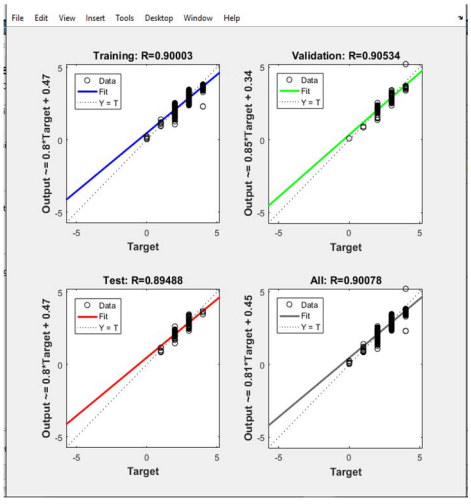

Fig 11: Regression analysis using ANN with all attributes

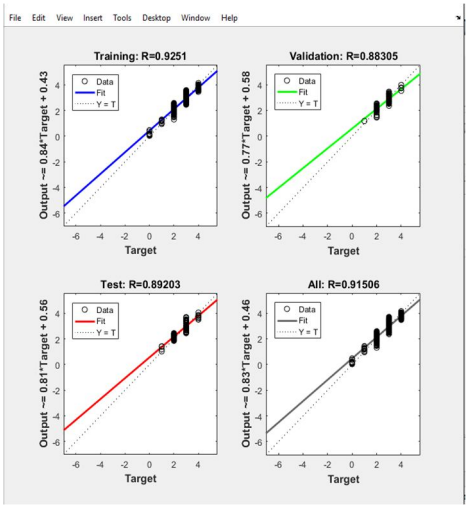

Fig 12: Regression Analysis using ANN after removal of nitrite

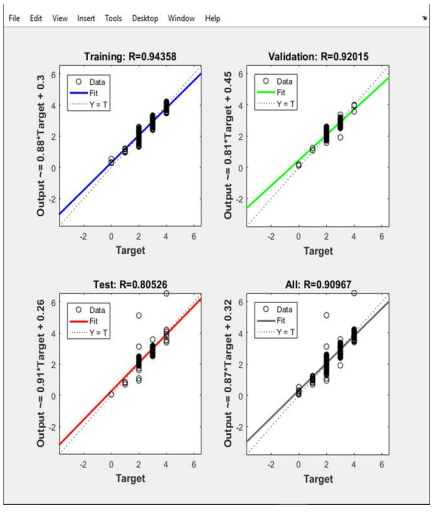

Fig 14: Regression Analysis using ANN after removal of turbidity attribute

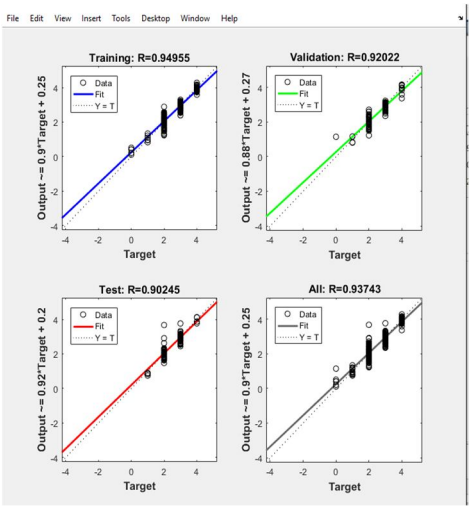

Fig 13: Regression Analysis using ANN after removal of ammonium attribute

As there is drop off of regression value in ANN after removal of turbidity attribute so only last 2 attributes will be reduced i.e ammonium and nitrite which results in increase of regression value in ANN. So, this reduces the variables from 15 to 1Further, Classification using decision trees with all the attributes is shown in fig l5. The accuracy is computed as 96.7 by considering all the attributes.
Classification using decision tree after removal of the lowest ranking attribute nitrite is shown in fig 16 and the accuracy is computed as 96.8.
Classification using decision tree after removal of the second lowest ranking attribute ammonium is shown in fig 17 and the accuracy is computed as 96.9.
Classification using decision tree after removal of the third lowest ranking attribute turbidity is shown in fig 18 and the accuracy is computed as 93.3.

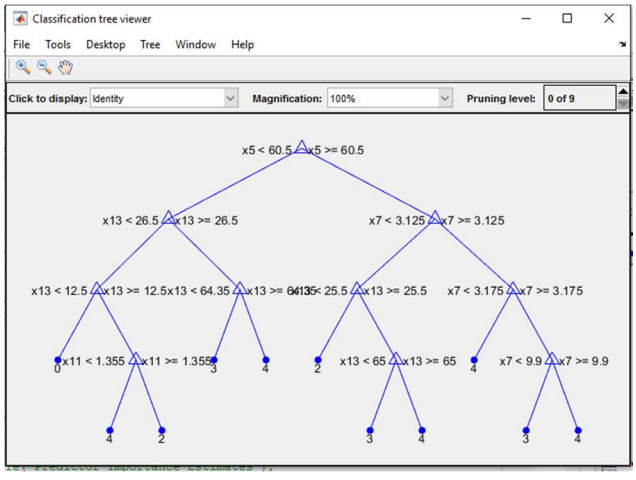

Fig l5

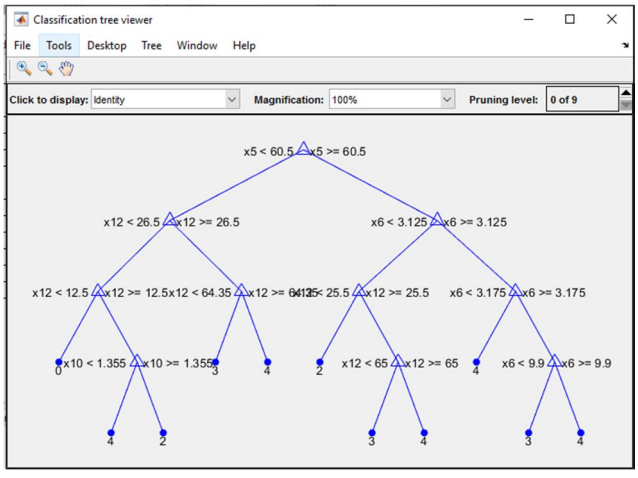

Fig 16

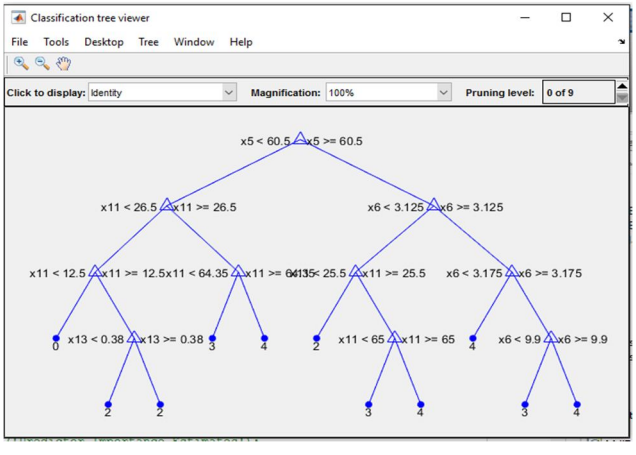

Fig 17

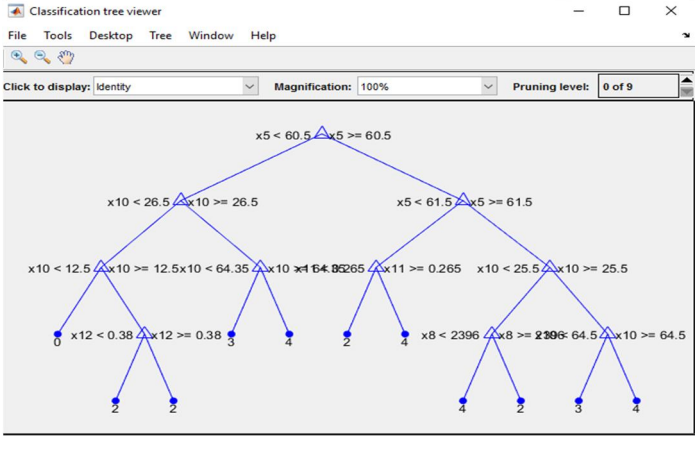

Fig 18

As there is drop off of accuracy in DECISION TREE classifier after removal
of turbidity attribute so only last 2 attributes will be reduced. i.e nitrite and ammonium which results in increase of regression value in DECISION TREE.
So, in this case also, the variables are reduced from 15 to 13.
Comparison of accuracy of ANN and DECISION TREE classifier is shown in Table3

Table 3

ANN	DECISION TREE
93%	96.9%

The result of the third part of approach is given through correlation plot between weather and water dataset as shown in Fig 20.

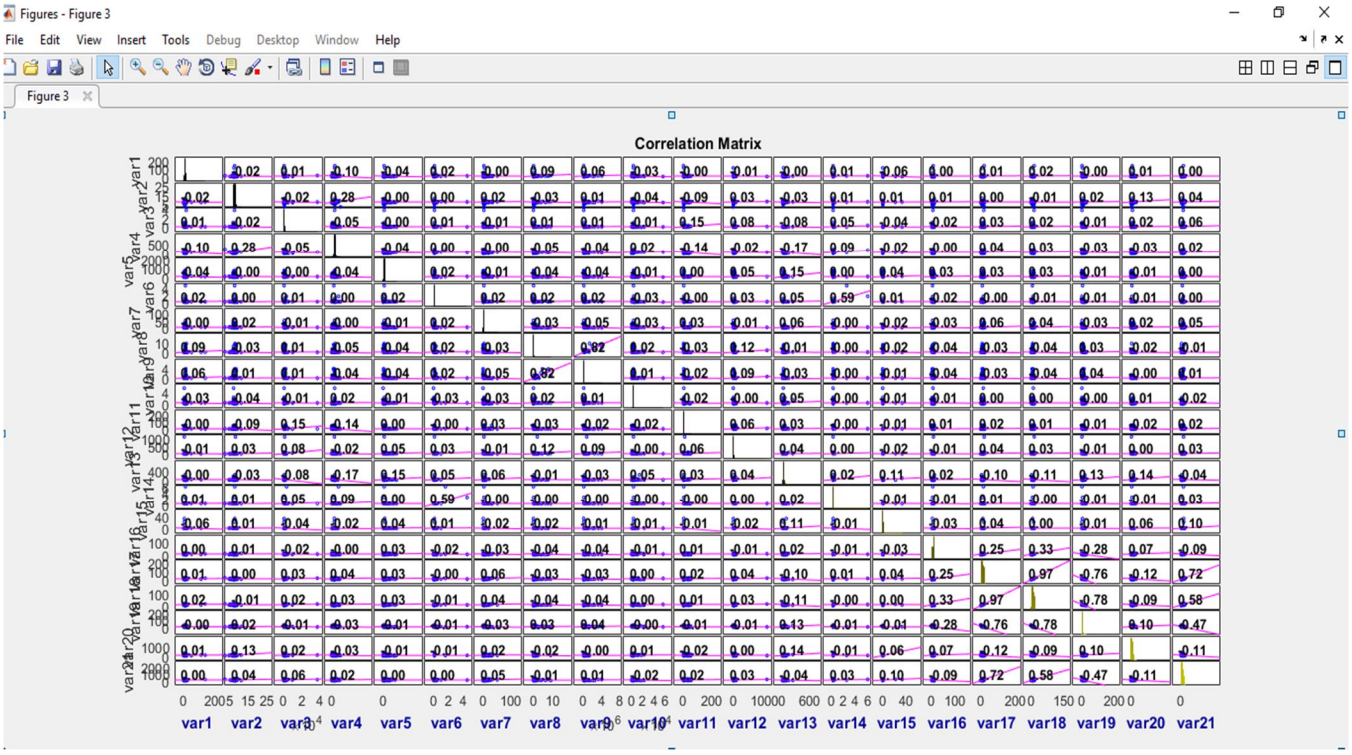

Fig 20

The attributes with values greater than 0.50 are considered as correlated. The attributes are either positively correlated or negative correlated. The plot shows the interdependence between the two datasets.
The next step results are shown by implementing ANN with NARX model.The performance measures of this model is given by Regression analysis.

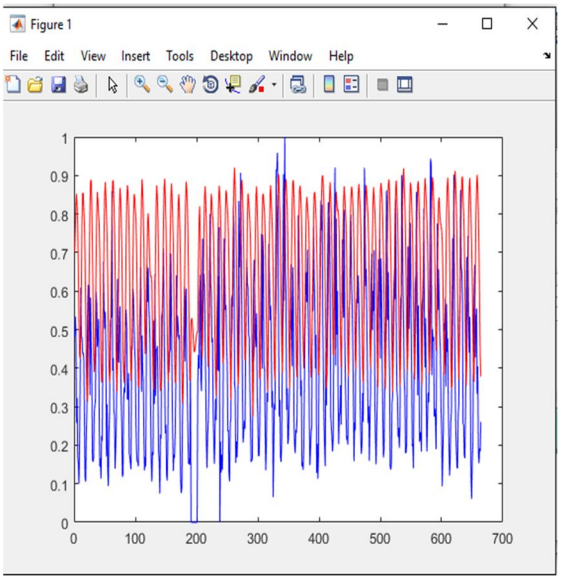

Fig 21:Prediction of co-related attribute Tmin and sun hours using NARX model

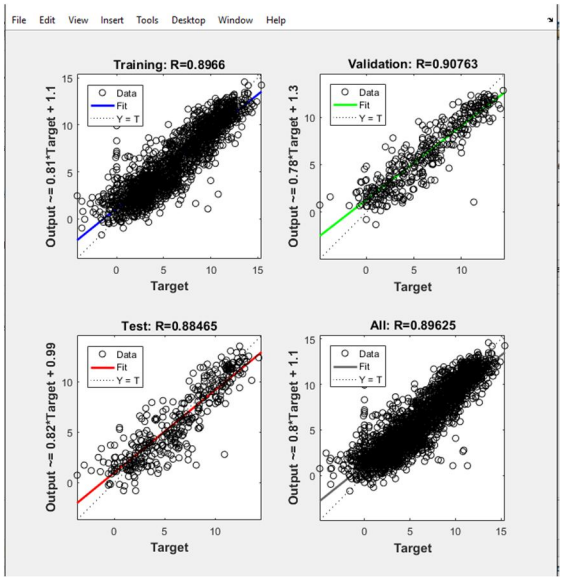

Fig 22: Regression Analysis

This graph shows how well the NARX model fits the function for training, validation and testing.

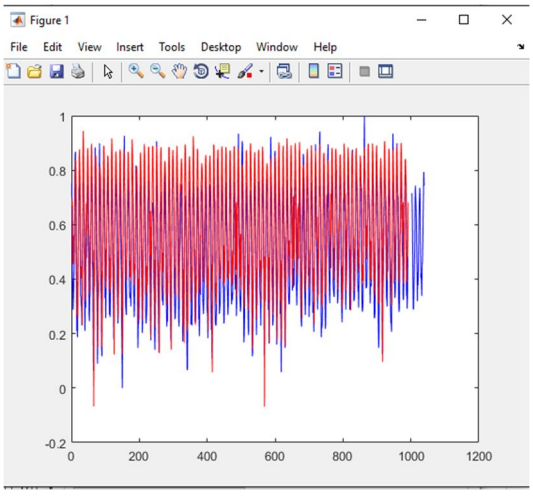

Fig 23: Prediction of co-related attribute C.O.D and WQI using NARX model

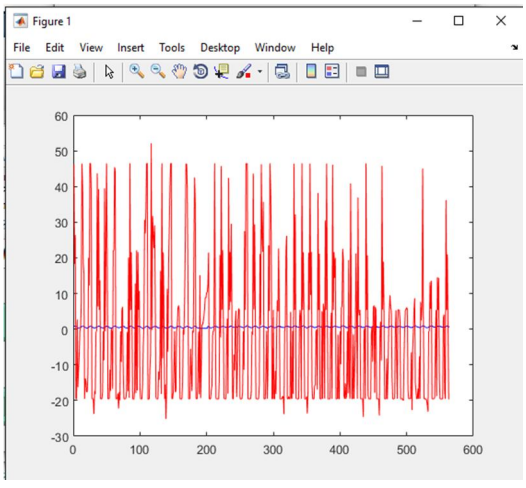

Fig 25: Prediction of co-related attribute tmax and tmin using NARX model

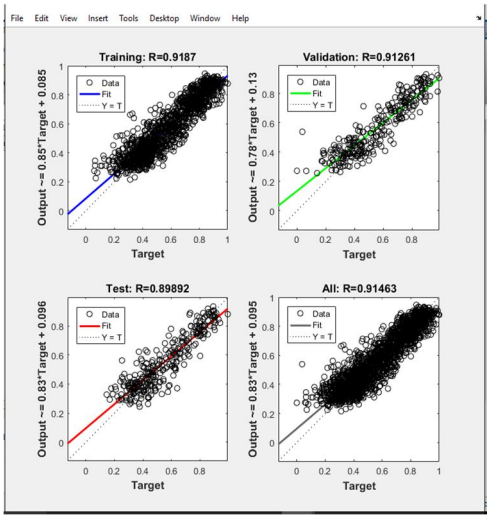

Fig 24: Regression Analysis

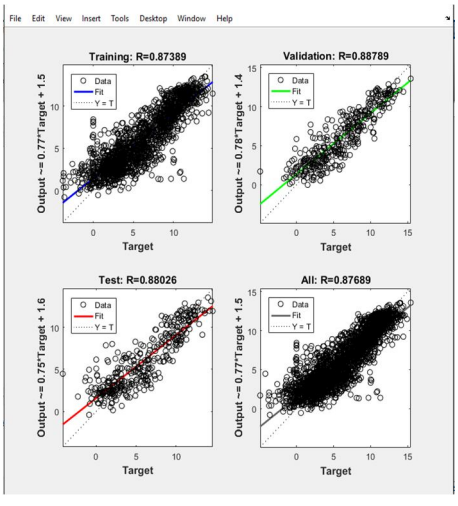

Fig 26: Regression analysis

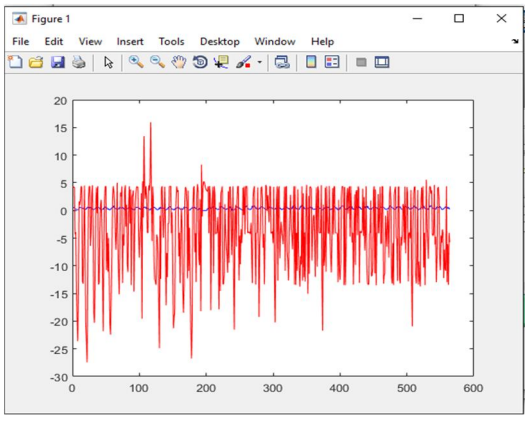

Fig 27: Prediction of co-related attribute tmin and air frost using narx model

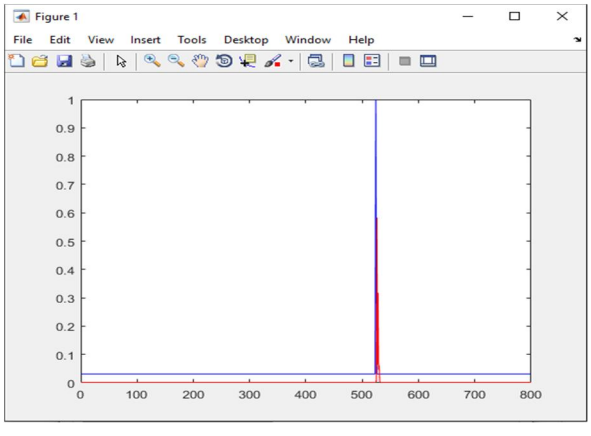

Fig 29: Prediction of co-related variables b.o.d and do using narx model

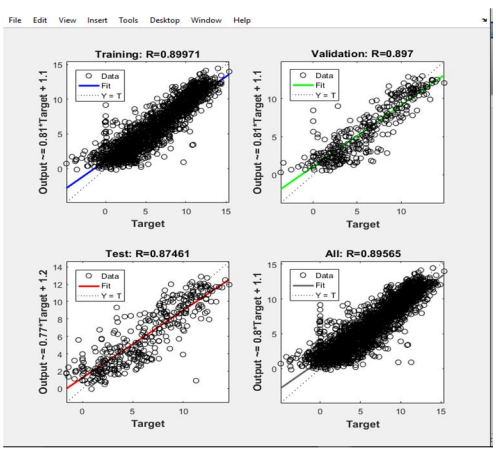

Fig 28 : Regression analysis

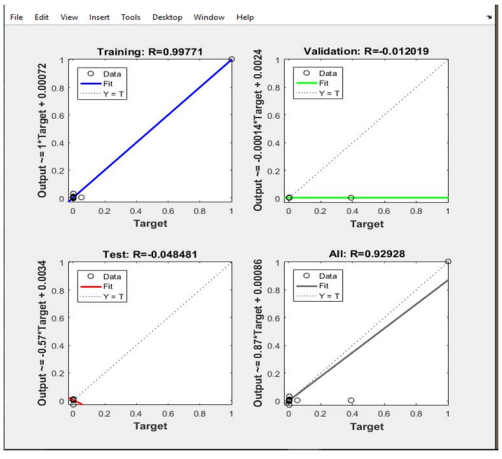

Fig 30: Regression Analysis

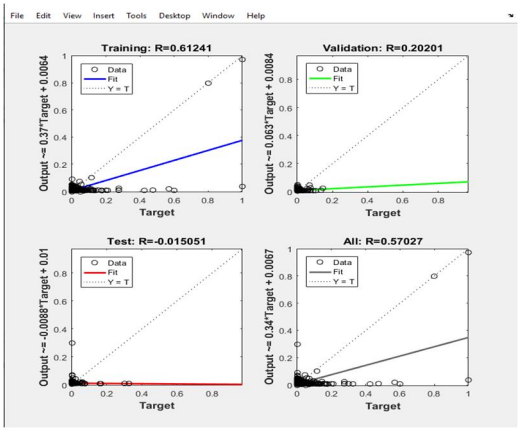

Fig 31: Regression Analysis

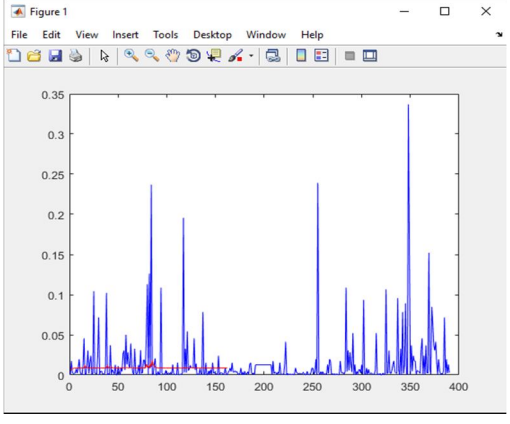

Fig 32: Prediction of co-related variables b.o.d and do using narx model

This approach has successfully predicted the future water quality trends by predicting the values for attributes which affect the quality of water by using ANN with NAR MODEL and NARX MODEL for co-related attributes. Such analysis of water quality would help to take some necessary actions to improve the quality of water. This approach has also identified the relevant attributes and reduce the non-relevant attributes from the data set. The impact of attribute reduction have been studied by analyzing the results thus obtained.

Table 4: Proposed Approach Results

Algorithms used	Results
ANN with NAR time series model	0.0006 MSE for phosphate and 0.8293 Regression value for phosphate
Dimensionality reduction	13 attributes
classification	93% for ANN and 96.8 for DECISION TREE
NARX MODEL	0.92928 regression value for correlated attribute B.O.D and DO

Acknowledgement

I would like to sincerely thank my institute and department for constant support and guidance as and when required for completing the work. I would also like to thank all the members of the department for their support and suggestions while completing the work.

References

[1] Yafra khan and Chai Soo see, "Predicting and analyzing water quality using Machine Learning: A comprehensive model", IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2016

[2] Jan-Yee-Lee, "Applying theory in predicting the arsenic contamination of groundwater in historical blackfoot disease territory", IEEE ninth international conference on natural computation,2013

[3] Maqbool Ali, Ali Mustafa Qamar ,"Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed in Pakistan",2013

[4] Ivana D. Radojević, Dušan M. Stefanović, Ljiljana R. Čomić, Aleksandar M. Ostojić, Marina D. Topuzović1 and Nenad D. Stefanović, "Total coliforms and data mining as a tool in water quality monitoring", African Journal of Microbiology Research Vol. 6(10), pp. 2346-2356, 16 March, 2012

[5] Neetu Arora, Amarpreet Singh Arora, Siddhartha Sharma and Dr. Akepati S. Reddy,'"Use of Cluster Analysis-A Data mining tool for improved water quality monitoring of river Satluj", International Journal of Advanced Networking Applications (IJANA), 0975-0290

[6] Ting-Nien Wu and Chiu Sheng Su, " Application of Principal Component Analysis and Clustering to spatial location of Groundwater Contamination", IEEE Fifth international conference on fuzzy systems and knowledge discovery,2008

[7] Jiawei Han and Micheline Kamber, "Data Mining : Concepts and Techniques", 2nd edition