Prediction of Trophic State Index of Lentic Water Bodies Using Artificial Intelligence- A Case Study

Arya S Nair
M Tech Student, Environmental Engineering and Management,
UKF College of Engineering and Technology,
Kollam, Kerala, India
sreearya818@gmail.com

Sruthy G S
M Tech Student, Environmental Engineering and Management,
UKF College of Engineering and Technology,
Kollam, Kerala, India
sruthysreekantan009@gmail.com

Preeja Prameelan
Assistant Professor, Department of Civil Engineering,
UKF College of Engineering and Technology,
Kollam, Kerala, India
preejapi@gmail.com

Adarsh S
Associate Professor, Department of Civil Engineering,
TKM College of Engineering
Kollam, Kerala, India
adarsh_lce@yahoo.co.in

Priya K L
Assistant Professor, Department of Civil Engineering,
TKM College of Engineering
Kollam, Kerala, India
klpriyaram@gmail.com

***Abstract*—Eutrophication of lentic water bodies have significant impact on the natural environment and human life. The quantities of nitrogen, phosphorous and other biologically useful nutrients are the primary determinants of a water body's Trophic State Index (TSI). The artificial intelligence (AI) methods have enhanced the estimation of Eutrophication status of lentic systems like fresh water lakes. This study uses three artificial intelligence methods such as M5 Model Tree, Support Vector Machine (SVM) for the prediction of Trophic states of Sasthamcotta fresh water Lake, Kerala. Four different cases were considered in the study via prediction of individual trpohic status (TSI-Ph, TSI-Ch1-a and TSI-Sechi Depth) along with the classical Carlson's Trophic State Index (CTSI). Several statistical measures were used to quantify the performance of the three AI methods. Both the tree based algorithms were found to be successful in accurate prediction of TSI with Random Forest method as the best among them.**

Keywords—Artificial intelligence, data mining, trophic state index, water quality parameters.

I. INTRODUCTION

Eutrophication refers to the nutrient enrichment of water bodies particularly with nitrogen and phosphorus compounds. It is considered and is considered as one of the serious ecological problems related to lentic water bodies such as ponds, reservoirs, lakes [1-3]. United Nations Environmental Protection (UNEP) reported that globally 30-40 % of lakes and reservoirs show tendency toward varying degrees of eutrophication. It severely deteriorates water quality leading to increased turbidity, loss of biodiversity, health hazards, diminishing aquatic growth caused by depletion of oxygen, and foul taste and odour. This, in turn, creates socioeconomic challenges, such as increased water treatment costs, difficulties in fulfilling the criteria for disinfection by-products, and aesthetic damage [4].

The conceptual understanding of eutrophication will be helpful in classification of water bodies into different trophic states and it is an essential step in management of fresh water ecosystems. In the past, both univariate and multivariate statistical approaches have been applied in the past for classification of trophic states. The univariate approach for trophic state classification is based on specified ranges of the cause (Nitrogen, Phosphorus) or response (Chlorophyll-a[Chl-a] and Secchi Depth [SD]) variables or on the variable information expressed in the form of indices. Sawyer [5] and Vollenweider [6] classified trophic states purely on the basis of nutrient values (such as Phosphorous and Nitrogen) while later on the variables such as water clarity and Chlorophyll parameters were included in the uni-variate approaches [7].

Carlson Trophic State Index (CTSI) [8] is perhaps the most popular index for determination of lentic water bodies. The freshwater bodies have been classified into three possible trophic states- *oligotrophic*, *mesotrophic* and *eutrophic* based on the level of eutrophication in freshwater bodies. The advancements of Artificial Intelligence (AI) techniques helped in the improvements in estimation of TSI of water bodies. In the past, the most popular Artificial Neural Network (ANN) or other data mining methods have been used for prediction of eutrophication of lentic water bodies like lakes or reservoirs. Many of the studies estimated the TSI as the indicator for eutrophication while some studies used Chl-a or algal bloom as indicator. For example, Jan-Tai Kuoa [9] used ANN for modeling eutrophication process of Te-Chi Dam reservoir on Taiwan. They simulated dissolved oxygen (DO), total phosphorus (TP), chlorophyll-a (Chl-a), and secchi disk depth (SD) and replaced the results in the Carlson linear equation. Huo et al., [10] predicted the above variables of Lake Fuxian in southwestern China, considering appropriate water quality parameters as inputs. Finally, the Carlson index was estimated as eutrophication state of dam reservoir. Saghi et al. [11] used ANN for prediction of TSI in the Dez Dam reservoir considering Chl a, SD, DO as the input data for modeling TSI. Kim et al. [12] predicted Chl-a concentration in Daechung reservoir, Korea using ANN. They considered COD, T-P, and PO4-PPO4-P as inputs and

reported that data normalization will significantly influence the results. Li et al. [13] used two machine learning methods, including support vector machine for regression (SVR) and random forest (RF) for prediction of Chl-a concentration of Baiyangdian Lake using a multitude of input variables. To improve the model accuracy and reduce the input number, two feature selection methods, including minimum redundancy and maximum relevance method (mRMR) and RF, were integrated with regression models. The results showed that the RF model had a higher predictive ability than the SVR model. Chou et al. [14] compared the performance of a number of machine learning approaches for predicting Carlson's Trophic State Index (CTSI), of reservoir in Taiwan. They used ANNs, SVM, RF, FNN and M5 Model Tree, have been used to analyze different scenarios. It was concluded that ANN displayed best performance in predict CTSI. Aria et al. [15] used a multilayer perceptron (MLP) and time delay neural network (TDNN) were used to predict the eutrophication status of two monitoring stations in the Amirkabir Reservoir in Iran. Six scenarios for each monitoring station were performed to select a significant, independent input using 12 years of monthly data. Temperature, turbidity, phosphate (PO4), nitrate (NO3), nitrite (NO2), ammonium (NH3), dissolved oxygen (DO) and electrical conductivity (EC) were the inputs considered and TDNN was reported to be performing better than MLP. Sensitivity analysis of the Amirkabir Reservoir dataset indicated increasing the value of nitrate is the first factor, followed by turbidity and NH3, having the greatest impacts on eutrophication prediction.

In Indian context, some attempts were made in estimation of Eutrophication status of lakes using techniques including ANN [16-18]. But no major study was reported in the prediction of TSI of Ramsar sites of Kerala using the advanced data driven methods. This study compare the performance of three sata driven methods MTree, SVM and RF in the prediction of TSI and the CTSI of Sasthamcotta Lake southern Kerala..

II. METHODOLOGY

In this study three AI methods are used for the prediction of TSI. The algorithmic background of the three methods are presented in brief in the following

A. M5' Model tree

M5 Model Tree (MT) proposed by Quinlan [19] is a popular machine learning technique used for solving regression problems through classification and decision making. Model Trees follows a modular approach so that the entire domain is divided into sub-domains and multi-linear regression models are developed for each sub-domain. Therefore it formulates many piecewise linear models to approximate the non-linear relationship between the input variables and output variable. In the first stage in the development of MT model, a decision tree is created following a splitting criterion. Depending upon the method of splitting of the domain, there exist different learning algorithms for model trees. The one which uses standard deviation reduction (SDR) as the splitting criteria is called as M5 learning algorithm [20]. In this method the standard deviation of the class values that reach a node is treated as a measure of the error at that node and the expected reduction in this error as a result of testing each attribute at that node is calculated. The computation of SDR can be represented as follows:

$$\sigma_R = \sigma(N) - \sum \frac{|T_i|}{|N|}\sigma(T_i)$$ where, σ_R= standard deviation reduction; N is the total number of training samples; T_i is the training samples of i^{th} sub-domain; $\sigma(N)$ and $\sigma(T_i)$ are the standard deviations of complete training samples, and i^{th} sub-domain samples respectively. A typical representation of this model could be in the form: $O = a_0 + a_1x_1 + a_2x_2 + ...$ where, O is the output values of the sub-domain; a_0, a_1, a_2…. are the coefficients of linear regression; x_1, x_2… are the input values of the sub-domain.

The splitting continues till the class values of all the instances that reach a node varies negligibly or only a few instances remain. Then the model improvement is done by two ways -performing 'pruning' and 'smoothing' operations, which reduce the effect of 'overfitting' and sharp discontinuities between different sub-classes, which happen especially when the dataset is very small [21]. A more detailed description of the theory behind model trees and pictorial representations can be found elsewhere [20].

B. Support Vector Machine

SVM (Support Vector Machine) proposed by Corts and Vapnik [22] is a supervised machine learning algorithm, which can be used for the classification and regression problems. SVM performs the regression by using a set of non-linear functions called Kernel functions that are defined in a high dimensional space. SVM has been used to solve non-linear regression problems by structural risk minimization (SRM) [23], in which the risk is measured using Vapnik's accuracy intensive loss function (ε) The SV regression constructs a hyperplane that minimizes the sum of the distances from the data points to the hyperplane. It is used to construct an input-output model for solving non-linear regression problems. During training, SVR kernel functions (linear, radial basis, polynomial, or sigmoid) are used to identify support vectors along the function surface [16].

C. Random Forest

The random forest (RF) method developed by Breiman [24] is an ensemble learning technique. It has been successfully used in dealing with various prediction problems. It is a machine-learning algorithm that combines a large set of decision trees to improve the prediction performance of classification and regression trees (CART) method. Each decision tree of RF is grown by using a randomly selected bootstrap sample from the original data set, and the final outcome of RF is the average result of all the trees. Compared to the regression methods, the number of parameters needed to be defined in the RF is very few. There are only two necessary parameters, including the number of variables used in each tree-building process and the number of trees built in the forest. The number of trees built in the forest has significant influence on the result of

RF. The insufficient number of trees would result in poor forecasting performance, while the excessive number of trees may lead to complicated predictors. WEKA (Waikato Environment for Knowledge Analysis) developed at the University of Waikato in New Zealand, is used for the implementation of different algorithms. The performance evaluation was done based on the most popular statistical measures such as root mean square error (RMSE), coefficient of correlation (R) and mean absolute error (MAE).

III. STUDY AREA AND DATABASE USED FOR DEVELOPMENT OF MODELS

The study area is Sasthamcotta Lake located in Kollam District, Kerala. The lake is located physiographically in the midland region between 90 0' - 90 5' N latitude and 760 35' - 760 46' E longitudes at an elevation of 33m above MSL. The lake has a catchment area of 12.69 Sq. km, surface area of 373 hectares, average depth of 6.53 m, maximum depth of 15.2m and a storage capacity of 22.4 Cu. km. This area is selected for study because this lake is a drinking water source for over 700000 people in Kollam district. An alarming fall in water level and pollution has put the biggest fresh water lake in Kerala at risk. According to Wetland Conservation and Management Rules, 2010 by the Ministry of Environment, Forest and Climate Change, the water bodies listed under the Ramsar convention centre are not to be polluted or encroached upon. Figure 1 shows the map of Sasthamcotta Lake.

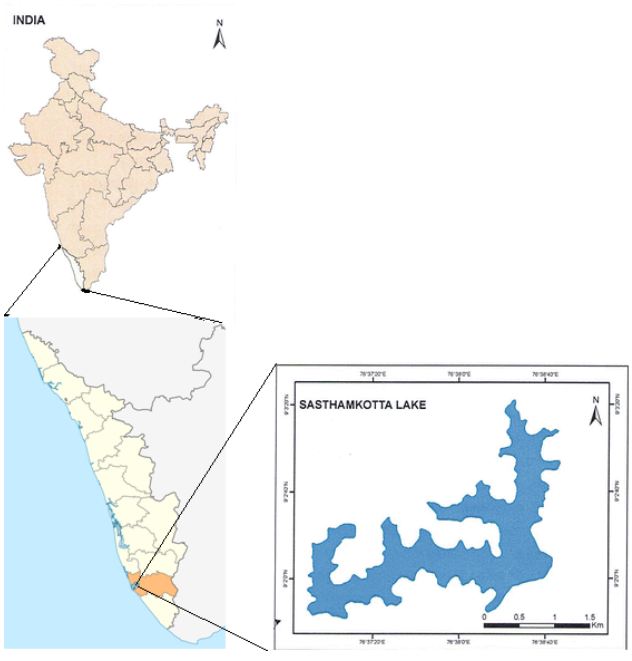


Fig. 1.Location of Sasthamcotta Lake

The data base consisting of 84 data points collected between 2012-18 were used for the study, which are used for calculating TSI. The values of parameters of the water quality model are collected from the Pollution Control Board of Kerala (PCB). The input parameters considered are Temperature, Electrical Conductivity, Total Nitrogen, Chlorophyll a, Total phosphorus, Seechi Depth. Statistical properties of the data are provided in the Table 1. The algorithms gives best result in 75% train ratio and 25% test ratio.

Table 1: Statistical properties of parameters. STD stands for standard deviation

Data	Property	T	EC	TN	Chla	TP	SD
Training	Maximum	30.5	190	16790	24.9	5.372	3.108
	Minimum	23.6	48	440	10.2	1.498	0.252
	STD	1.357	26.26	2135.25	2.983	0.55	0.623
	Mean	25.448	93.657	2224.25	15.696	2.999	0.994
Testing	Maximum	28	164	1980	19.8	3.095	3.092
	Minimum	22	40	590	11.5	1.784	0.768
	STD	1.618	27.645	349.184	2.559	0.361	0.472
	Mean	25.5	123.16	916.667	15.056	2.165	1.359

IV. RESULTS AND DISCUSSIONS

In the prediction of TSI, firstly, the TSI of individual variables and CTSI were computed using the following formulae [25]

TSI (SD) = 60 – 14.4ln(SD)

TSI (Chl) = 9.81 ln (Chl) + 30.6

TSI (TP) = 14.42 ln (TP) + 4.15

Equation for CTSI

CTSI= (TSI(SD)+TSI(Chl)+TSI(SD)) / 3

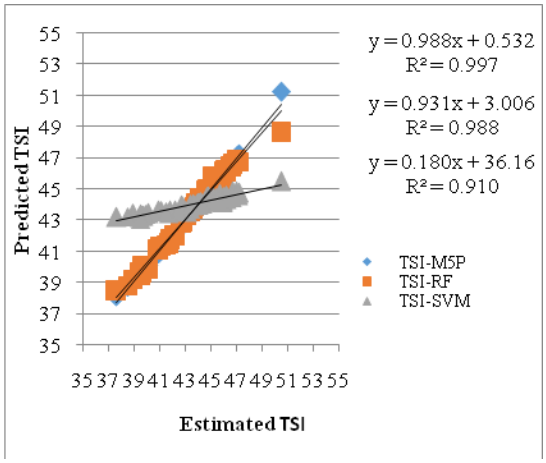
Measuring the concentration of Chlorophyll-a and Total phosphorus requires sophisticated techniques that involve chemical experiments, equations, specific analyses of water samples. In the absence of field Chl-a data, it can be determined from easily measurable water quality parameters as follows [26]

log Chl a = -3.04-0.25 log NO3 +0.58 log NO2 +0.45 log TN/TP+ 1.582 log Temp+0.65 log EC

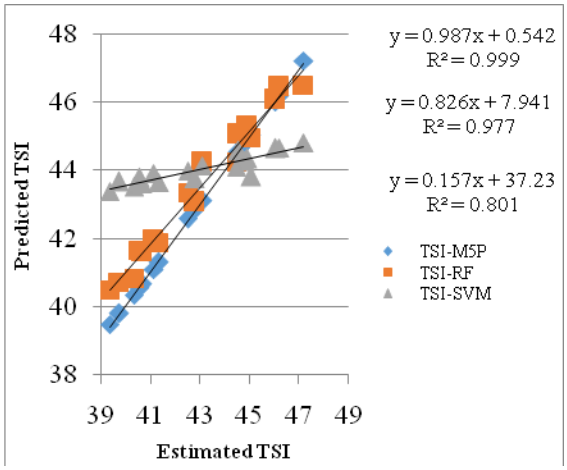
The determination of TSI-TP, TSI-N, TSI-Chl-a and CTSI are referred as Cases-1-4 in this paper. The data base was trained and tested for different train to test ratios. Finally a ratio of 75% train to 25% test was observed to give sufficiently good results for the three methods adopted for comparison. In the implementation of M5 model tree, no special parameter setting is required and a pruned and smoothed version is followed to control overfitting The developed model has three instance and is based on if -then conditions. In the SVM, different Kernel functions were tested and results were found to be better when we use RBF kernel. The hyper-parameter C is changed which is used to determine the trade-off between the complexity of the decision rule and the frequency of error, and that was kept to 0.1. SMOreg algorithm was used for the optimization of parameters. The RF model was developed with TSI as the output parameter. Bagging with 100 iterations was performed with the bag size (as a percentage of the training set size) fixed at 100. It is to be noted than in the

development of CTSI, all the three variables TP, TN and Chl-a were considered as inputs.

The models were compared on the basis of error measures and the scatter plots between predicted Vs estimated values of TSI. Table 2 shows the performance evaluation of different models for different cases. Figures 2-5 presents the scatter plots of prediction for the four cases.

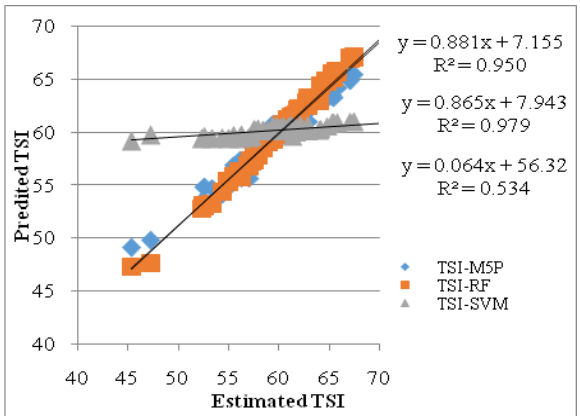


a.Train data

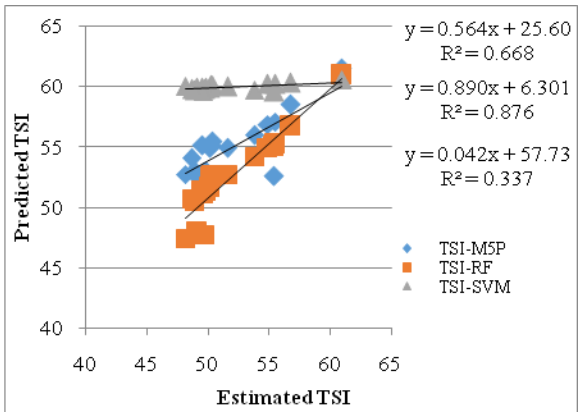


b.Test data

Fig. 2 Scatter plots of predictions of TSI-TP for training and testing (a) training data (b) testing data

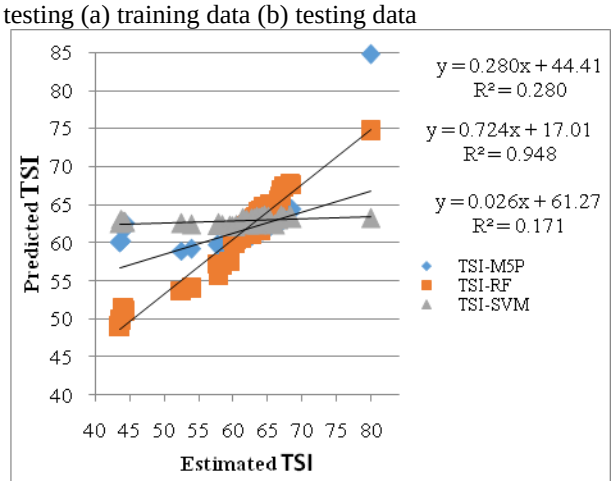


a. Train data

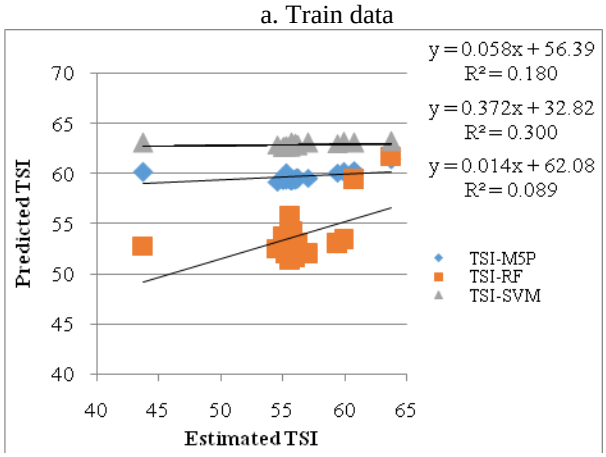


b. Test Data

Fig. 3 Scatter plots of predictions of TSI-TN for training and testing (a) training data (b) testing data

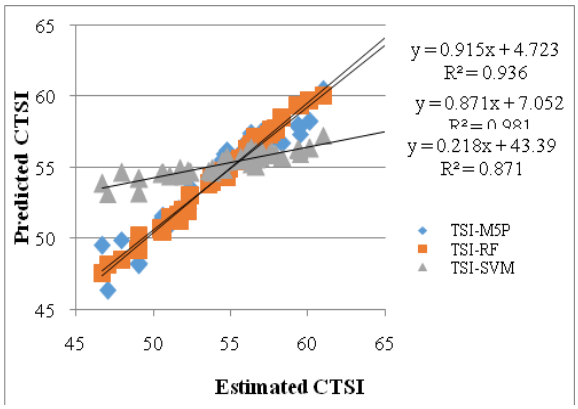


a. Train data

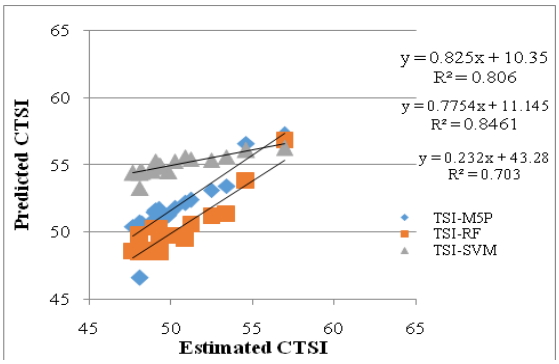


a. Test data

Fig. 4 Scatter plots of predictions of TSI-Chl-a for training and testing (a) training data (b) testing data



a. Train data



b. Test data

Fig. 5 Scatter plots of predictions of CTSI for training and testing (a) training data (b) testing data

.

Table 2: Performance evaluation of TSI using different methods.

Model	Training set				Testing set			
	R	RMSE	RAE	MAE	R	RMSE	RAE	MAE
Case -1								
M5P	0.998	0.138	3.450 %	0.079	0.999	0.052	2.010 %	0.043
SVM	0.954	2.324	79.561 %	1.827	0.895	2.239	87.188 %	1.887
RF	0.994	0.339	8.623 %	0.199	0.988	0.690	27.141 %	0.587
Case -2								
M5P	0.975	1.246	28.081 %	0.987	0.817	3.666	40.655 %	3.369
SVM	0.731	5.017	91.618 %	3.222	0.581	8.746	98.287 %	8.145
RF	0.989	0.996	9.667 %	0.34	0.936	1.362	13.682 %	1.133
Case -3								
M5P	0.53	5.324	84.294 %	3.403	0.425	5.088	66.263 %	3.857
SVM	0.406	6.227	87.569 %	3.536	0.300	7.728	116.79 %	6.799
RF	0.973	2.031	26.455 %	1.068	0.771	3.556	60.706 %	3.098
Case -4								
M5P	0.967	0.970	26.318 %	0.741	0.897	1.883	34.403 %	1.710
SVM	0.933	3.034	76.614 %	2.159	0.838	5.019	94.846 %	4.716
RF	0.990	0.674	11.658 %	0.328	0.919	0.992	16.435 %	0.817

From Table 2 it is noted that case 1 and case 4 gives much better results than case 2 and case 3 by different methods. In case 1 and case 4, R value for prediction by all the three methods is greater than 0.9 and training was fund to be greater than 0.8. In case 2, the two tree based methods display good accuracy in prediction (R value of 0.81 and 0.94; RMSE of 2.66 and 1.36 respectively by M5 model tree and RF). In case-3, except random forest, the other methods fails to capture the relationship even in the training phase, This may be because of the fact that instead of accurate real field observation, a proxy equation was used in the estimation of Chl-a. On examining the results of different cases by the three AI methods, it is noted that RF method consistently performs well (highest R value and lesser error measures) in all scenarios. In case-1, M5 model tree was found to be competent and marginally better than RF. The CTSI, which is a more realistic representation of trophic state, the RF shows high R value (0.92 against 0.89 and 0.84 by M5 model tree and SVM) and less RMSE (0.99 by RF against 1.88 and 5.01 by M5 model tree and SVM) in the testing phase. Thus both the uni and multi-variate associations in the estimation of trophic states could be captured well by the AI methods. The better performance is noted by tree based modeling than the purely non-linear SVM model. A rigorous parameter optimization by integrative soft computing methods may slightly improve the results of SVM, which need to be verified by soliciting more experiments.

V. CONCLUSION

The study developed AI models for the prediction of TSI of Sasthamcotta Lake employing SVM, M5 Model tree and Random Forest methods. Four cases were considered including the predictions of individual TSI and the CTSI. Both the tree based AI methods found to be effective in prediction of TSI and RF method is performing consistently well in all the four cases. This study contributes to the improvement of water quality management of lentic water bodies by presenting a set of versatile techniques that offers diverse predictive capabilities.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Dr. Adarsh S, Associate Professor, TKMCE, Kollam, Kerala, for providing me enough information and guidance to complete this work successfully. I also express my gratitude to all those who helped me to perform this work.

REFERENCES

[1] W. Xiao-long, Y.L. Lu, G.Z. He, J.Y. Han, and T.Y. Wang "Exploration of relationships between phytoplankton biomass and related environmental variables using multivariate statistic analysis in a eutrophic shallow lake: a 5-year study. *Journal of Environmental Science* Vol. 19, pp.920–927 2007

[2] L. Ye, Q. Cai, M. Zhang, and L. Tan, "Real-time observation, early warning and forecasting phytoplankton blooms by integrating in situ automated online sondes and hybrid evolutionary algorithms." *Ecological Informatics* Vol. 22, pp. 44-51, 2014

[3] Cao, C., Zheng, B., Chen, Z., Huang, M., and Zhang, J., 2011. Eutrophication and algal blooms in channel type reservoirs: a novel enclosure experiment by changing light intensity *Journal of Environmental Science* Vol. 23, pp. 1660–1670, 2011

[4] V. H. Smith "Eutrophication of freshwater and coastal marine ecosystems:a global problem" *Environmental Science Pollution Research Int*. 10,126–139, 2003

[5] C.N. Sawyer "Fertilization of lakes by agricultural and urban drainage" *Journal of New Englland Waterworks Association*. Vol. 61, pp.109–127, 1947

[6] R.A. Vollenweider *The scientific basis of lake and stream eutrophication, with particular reference to phosphorus and nitrogen as eutrophication factors*. Technical Report OAS/DSI/68.27. Organization for Economic Cooperation and Development. Paris., 1968

[7] C. Forsberg, and S. Riding, "Eutrophication parameters and trophic state indices in 30 Swedish waste-receiving lakes" *Arch. Hydrobiol.* Vol. 89, pp. 189–207, 1980

[8] R. E. Carlson, "A Trophic State Index for Lakes," *Limnology and Oceanography*, Vol. 22, pp. 361-369, 1977

[9] J-T Kuo, Ming-Han Hsieh, Wu-Seng Lung and Nian She "Using artificial neural network for reservoir eutrophication prediction" *Ecological modeling* Vol. 200,pp. 171-177, 2007

[10] S. Huo, Z He, J. Su, B. Zi, and C. Zhu "Using artificial neural network models for eutrophication prediction". *Procedia Environmental Sciences* Vol. 18, pp. 310 – 316, 2013

[11] H. Saghi, L. Karimi, and A. H. Javid, A. H, "Investigation on TSI by artificial neural network," *Applied Water Science*, pp. 127–136, 2015.

[12] J. Kim, J. Ki, and Y. Cao "Establishing a Predictive Model for Chlorophyll-A Concentration in Lake Daechung, Korea Using Multilinear Statistical Techniques" *Journal of Environmental Engg. ASCE* 141(2)

[13] X. Li, J. Sha and Z. Wang, "Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake," *Environmental Science and Pollution Research*, Vol. 25, pp. 488-498, 2018.

[14] J. S. Chou, Chia-Chun, Ho and Ha-Son Hoang, "Determining quality of water in reservoir using machine learning," *Ecological Informatics*, Vol. 17, 1574-9541, 2018.

[15] S.H. Aria, G Asadollahfardi and N Heidarzadeh (2019) Eutrophication modelling of Amirkabir Reservoir (Iran) using an artificial neural network approach. *Lakes & Reservoirs* Vol. 24, pp. 48–58, 2019

[16] A. G. D. Prasad and Siddaraju, "Carlson's Trophic State Index for the assessment of trophic status of two lakes in Mandya district," *Advances in Applied Science Research 3*, pp. 2992-2996, 2012.

[17] T Anthwal, A Chandola, M P Thapliyal. "Performance Analysis of ANN Model for Estimation of Trophic Status Index of Lakes". IAES International Journal of Artificial Intelligence (IJ-AI) 7(1):1-10, 2018

[18] A. M. Sheela, J. Letha, S Joseph "Environmental status of a tropical lake system". *Environmental Monitoring and Assessment* Vol. 180(1-4), pp. 427-449, 2011

[19] J.R. Quinlan, "Learning with continuous classes", in Proceedings of the Australian Joint Conference on Artificial Intelligence World Scientific, Singapore, 1992), pp. 343–348

[20] H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*, Morgan Kaufmann, San Mateo, 2000

[21]V. Jothiprakash, and A. Kote, "Effect of pruning and smoothing while using M5 model tree technique for reservoir inflow prediction" *Journal of Hydrologic Engineering* Vol. 16(7), pp. 563–574 , 2011

[22] C. Cortes, and V.N. Vapnik, V. N. "Support vector networks", *Machine learning*., Vol. 20, pp. 273-297, 1995

[23] V. N. Vapnik "*The nature of statistical learning theory*", Springer, New York (1995).

[24] L. Breiman "Random forests". Machine Learning 45(1), pp. 5–32, 2001

[25] R.E. Carlson, and J. Simpson "A Coordinator's Guide to Volunteer Lake Monitoring Methods", North American Lake Management Society.

[26] C.D. Stanley, R.A. Clarke, and B.W. MacLeod, SL-211, a publication of the Soil and Water Science Department, Florida Cooperative Extension Service, IFAS, University of Florida., 2003.