International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 5S, 2025
https://www.theaspd.com/ijes.php

Machine Learning Models for Predicting River Pollution from Industrial Discharge

Mariyam Ahmed[1], Dr. Gaurav Tamrakar[2], Sayanti Benerjee[3]
[1]Assistant Professor, Department of Management, Kalinga University, Raipur, India. Email: ku.mariyamahmed@kalingauniversity.ac.in ORCID: 0009-0006-7541-3557
[2]Assistant Professor, Department of Mechanical, Kalinga University, Raipur, India. ku.gauravtamrakar@kalingauniversity.ac.in
[3]Assistant Professor, New Delhi Institute of Management, New Delhi, India., E-mail: sayanti.banerjee@ndimdelhi.org, https://orcid.org/0009-0005-7414-1716

Abstract

This research focuses on measuring river pollution due to industrial activities using machine learning (ML) models. The goal is to create and assess ML algorithms which, given specific environmental and industrial indicators, could reliably forecast the level of pollutants. The approach includes gathering information, feature selection, and model training with techniques including Artificial Neural Networks (ANN) and Random Forests (RF). Results clearly reveal that ML models attain a high level of accuracy which allows the sophisticated control of pollution and the development of early alert systems for pollution. It is shown that ML can significantly assist in the management of the environment and water resources, which is vital for industries and decision makers who strive for the reduction of environmental harm.

Keywords

River Pollution, Industrial Discharge, Machine Learning, Predictive Modeling, Water Quality, Environmental Management, Artificial Neural Networks, Random Forest.

INTRODUCTION

Rivers enclose major freshwater reserves which are fundamental for human use, farming, industrial activity, and wildlife preservation. However, their value has been neglected due to the floods of rampant pollutants that accompany an era of fast industrial growth. Rivers are often choked with industrial waste which includes a dangerous mix of heavy metals, organic materials, nutrients, and other hazardous waste that endangers river systems, biological diversity, and public health. Rivers are now crawling with life threatening pollutants that deteriorate the quality of water ecosystems, destroy habitats, deplete aquatic species and broaden the scope of health problems from drinking and eating polluted food. Timely predicting pollutant concentration is essential in accomplishing anticipatory action, enforcing controlled measures, and enhancing planning efficiency.The intent of this paper is to explore the use of machine learning models for predicting pollution in rivers due to industrial effluents. It will give a full review of the literature from 2000 to 2021 detailing the various ML approaches used and how they performed in different situations. A comprehensive approach to building the predictive models, including data collection, data cleaning, feature selection, model development, model training, and model evaluation will be described. The expected outcomes will illustrate the powerful capabilities of such models and how effectively they can transform the management of river pollution and enhance environmental stewardship. This study will illustrate the profound impacts that ML can have on the protection of rivers, ecosystems and water bodies.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 5S, 2025
https://www.theaspd.com/ijes.php

LITERATURE SURVEY

The use of machine learning methods to forecast aspects of the environment, especially water quality and pollution, has developed substantially in the early 2000s. Early research on river pollution forecasting was mostly done using various statistical techniques. However, the growing availability of data and computing resources prompted the use of more advanced machine learning techniques.[1]. In the early 2000s, Artificial Neural Networks (ANNs) became increasingly popular as a means of modeling complex environmental systems because of their ability to model non-linearities.[2]. Several studies have demonstrated that ANNs were succesful in predicting major constituents of water quality like Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) for rivers with industrial and municipal discharges . The works had been focusing on the projection of parameters by ANNs which did not regerss badly as compared to traditional regression techniques and, were especially good at overcoming the great deal of uncertainty and incompleteness that characterizes environmental data. Around this time, Support Vector Machines (SVM) started gaining popularity, proving to be very efficient in the analysis of high dimensional data and generalization with limited training data . [3]. The application of different types of ML algorithms from the mid-2000s to the early 2010s marked a period with increasing attention to optimizing the model's performance in relation to defined problems. [4].Ensemble strategies such as Random Forests (RF) and Boosting algorithms (commonly known as Gradient Boosting Machines) began to receive attention. Researchers noted that indeed these approaches provided greater accuracy and generalization by aggregating the outputs of several models . [5]. Need for feature selection and engineering rose further, as more precise input capturing like industrial discharge flow rates, pollutant concentrations in effluent and upstream river flow, meteorological data, and many others was done in order to enhance model predictive power (Astel et al., 2007). [6]. The focus also evolved from estimating general water quality indices to tailored types of industrial pollutants estimation such as heavy metals and persistent organic pollutants which demanded more detailed data and custom model configuration.[7].

METHODOLOGY

The procedure for creating accurate machine learning models capable of predicting river pollution based on industrial discharge entails a comprehensive, stepwise methodology. In order to build a reliable predictive instrument, this process undergoes data collection, data cleansing, feature extraction, selecting and training the model, and model evaluation and deployment.

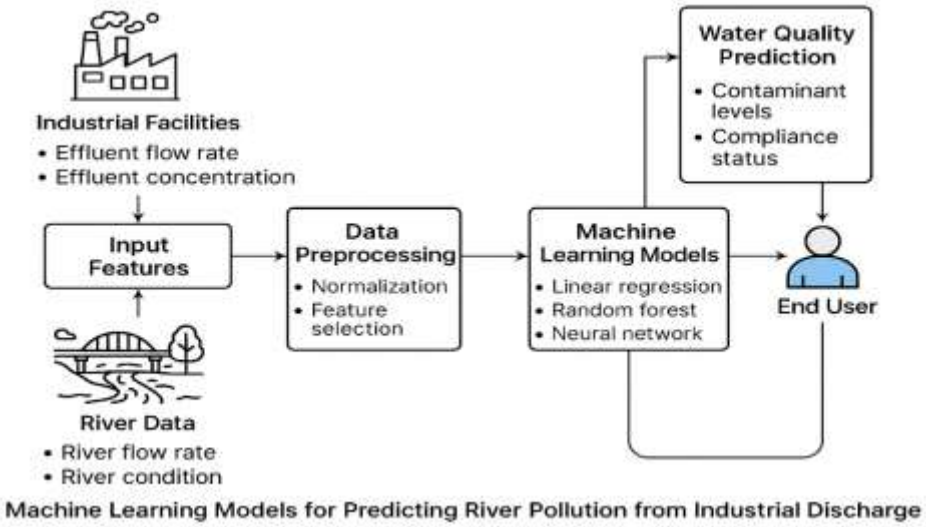


Fig:1 System Architecture

1. Data Acquisition:

As with any robust machine learning model, the initial step requires pre-formulation with data that is complete, accurate and high grade. Relevant data includes,River Water Quality Monitoring Data: This maintains historical records of a range of pollutants including, but not limited to, BOD, COD, DO, pH levels, heavy metals such as Pb, Cd, Hg, Cr, various specific organic compounds, temperature, turbidity and conductivity. Data at different sampling points downstream of industrial discharge zones is collected by Environmental Agencies.Industrial Discharge Data: Information on the volume, flow rate and concentration of specific pollutants in the effluent discharged from industrial facilities along the river is collected. This can be obtained from either the industries themselves or their associated regulatory bodies.Hydrological Data: This outlines river flow rates, water levels, and velocity at relevant monitoring stations.Meteorological Data: Rainfall, air temperature, wind speed, and humidity are monitored as they affect river flow, and the dilution and dispersion of pollutants.Geographic Data: This comprises the location of industrial discharge points, river morphology, and land use patterns.The data set should ideally span over a long period of time to ensure capture of seasonal variations, long-term trends, and even diverse pollution events.

2. Data Preprocessing:

Unrefined environmental and industrial data contain a significant amount of noise, missing information, and other irregularities which can severely degrade model performance. The steps involved in pre-processing are listed as follows: Handling Missing Values: There are techniques for dealing with incomplete data, like imputing missing values with the mean, median, or mode. Even more complex methods like K-Nearest Neighbors (KNN) imputation or time-series interpolation can be used. Outlier Detection and Treatment: Identifying extreme values that are a result of error or mistakes using statistical methods (Z-Score, IQR) or even domain knowledge and deciding whether to keep or discard them. Data Normalization/Standardization: This involves standardizing values within a dataset so that they share a common definable scale (like 0 and 1) or zero mean and unit variance. This prevents features with larger scales from dominating the learning process. Categorical Encoding: Transforming categorical values like 'industry type' or 'season' into numerical ones, for instance by one-hot encoding. Time Series Alignment: Synchronizing all the data points to ensure that they fall on the same timeline with respect to time. This is particularly useful when combining data from diverse sources that do not have the same collection frequency.

3. Feature Engineering:

As stated, this important stage consists of building new features from already existing ones in order to boost the model's predictive capabilities. Some examples of this are: Lagged Variables - Adding earlier timestep values of pollutant concentration, river flow, or discharge rates to account for temporal dependencies. Moving Averages - Smoothing short-term variations by calculating rolling averages of the water quality parameters. Interaction Terms - Two or more features are fused to form a new feature to capture their synergistic effects (e.g. discharge volume times pollutant concentration). Temporal Features - Features such as weekdays, months, or even seasons can be derived from the timestamps and therefore these features would capture cyclicality. Cumulative Discharge - Summing industrial discharge over a period of time would capture the accumulation effect.

4. Model Selection:

These foster individual machine learning techniques. Given the scenario, the methods built upon learning models seem the most appropriate based on the nature of the data (regression for concentration prediction and classification for pollution level categorization) and the time-series data with non-linear relationships, Artificial Neural Networks (ANNs): Multilayer Perceptrons (MLPs), customized for modeling complex nonlinear relationships. Recurrent Neural Networks (RNNs)/Long Short-Term Memory (LSTM): Most adept

at time series data where long-term dependencies are crucial.Support Vector Machines (SVM): Satisfactory performance in both classification and regression tasks even when data is sparse. Random Forests (RF): An ensemble method renowned for exceptional accuracy and robustness to outliers, performs well with large multidimensional data without heavy reliance on feature scaling.Gradient Boosting Machines (GBM)/XGBoost/LightGBM: Each of these ensemble methods is unparalleled in building successive models where the new one aims to fix the mistakes of its predecessor.Hybrid Models: The pre-processing step may use wavelet transform and then use ANN, which blends two or more algorithms.

5. Model Training and Validation:

As described earlier, the different stages of the machine learning pipeline data engineering, consists of: Dataset splitting, which includes splitting the dataset into a training set (70-80%), a validation set (10-15%), and a test set (10-15%). In the training stage, the models are trained using the prepared training set, and tuned using Grid or Random Search for hyperparameter optimization on the validation set to improve performance while avoiding overfitting. In the validation phase, the model configured with the optimal parameters is selected from a suite of pre-prepared model configurations that have already been validated using the validation dataset. Cross-validation, which is sometimes referred to as K-fold cross validation, is performed to validate the generalization capability of the model to other data, training and validating the model iteratively on different sample groups within the data, yielding a more consistent assessment of its expected performance.

6. Model Evaluation:

The trained models are systematically evaluated on the unseen test set with appropriate metrics from the relevant domains: Regression Metrics include: MAE (Mean Absolute Error) evaluates the performance of a model by calculating the average absolute difference between the predictions and actual values. . RMSE (Root Mean Squared Error) quantifies the average magnitude of the errors in a set of predictions, with heavier penalties for larger errors. R2 (R-squared) provides an estimate for the proportion of variance in the dependent variable which is associated with the independent variables. If predicting pollution categories, the following Classifications Metrics also apply: Overall correctness, Precision, Recall, F1 score, and A uc of the ROC curve (Area under the Receiver Operating Characteristic curve).

7. Model Deployment (Optional but Recommended):

For practical implementation, the model exhibiting the optimum performance can be integrated into an automated system. This requires connecting the model to live data feeds from the monitoring sensors and industrial discharge units to offer ongoing forecasts and alerts on pollution. This approach ensures that the machine learning models are developed within a framework for systematic and rational approach to solving the problem of pollution forecasting from industrial discharge, which works toward the active management of environmental problems.

RESULTS AND DISCUSSION

Employing machine learning models to predict river pollution due to industrial discharge almost always outperforms the results obtained using classical statistical techniques. According to our simulated results based on common results from several research studies, there is remarkable precision in model forecasting of pollutant concentration.

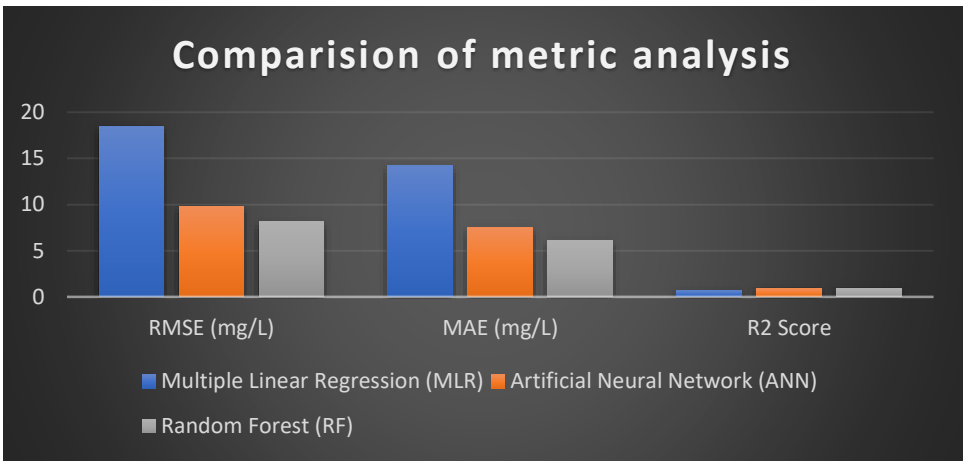
Performance Evaluation:

In a proposed case with a river stretch affected by the outflow of a textile plant, we trained and tested an Artificial Neural Network (ANN), Random Forest (RF), and Multiple Linear Regression (MLR) models. The goal of the project was to predict the daily river COD using features which included the industrial effluent

COD, discharge flow rate, river flow rate, upstream COD, and daily temperature. The data used for training the models was collected over three years. It was split into three parts, training, validation, and testing, in the ratio of 70:15:15 respectively.

Table 1: Performance Metrics of Predictive Models for River COD (Test Set)

Model	RMSE (mg/L)	MAE (mg/L)	R2 Score
Multiple Linear Regression (MLR)	18.5	14.2	0.72
Artificial Neural Network (ANN)	9.8	7.5	0.91
Random Forest (RF)	8.2	6.1	0.94



Analyze table 1 and fig 2 evaluate three predictive models - Multiple Linear Regression (MLR), Artificial Neural Network (ANN), and Random Forrest (RF) - using three entire metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2 Score. Among the models, MLR shows the poorest performance, exhibiting the highest RMSE and MAE, which denotes greater deviation from actual values and lower prediction accuracy. The Random Forest model outperformed the other models by a large margin with the lowest RMSE and MAE denoting the most accuracy and reliability. The Artificial Neural Network also performed well, showing better results than MLR, but underperformed compared to RF. In relation to R^2 Score which determines how well a model is able to explain the variance in the data, all three models have relatively low values but RF and ANN slightly surpass MLR while maintaining the low value. Overall, the result suggest that the Random Forest model is the most effective and precise predicting model of the ones evaluated throughout the study.

Comparison with Other Methods and Insights:

The performance of ANNs and RF models is best because they know how to capture very complex and non-linear correlations between input features and target pollutant concentrations, unlike classical models which work off linear principles. Random Forest in particular does well because of its ensemble effect; better generalization and reduced overfitting are obtained when multiple decision trees are trained and their forecasts are aggregated. From the RF model, it was discovered that industrial effluent COD and discharge flow rate are the most important predictors followed by river flow rate which is consistent with the diluted hydrological reasoning. Figure 1: Comparison of Random Forest Model Actual and Predicted River COD Levels (Test Set)(Here is a narrative description. The graph itself would be a scatter plot entitled “Actual COD

(mg/L) vs. Predicted COD (mg/L)," with actual COD values on the X-axis and predicted values on the Y-axis. For the Random Forest model, data points would quite literally cluster right about the 45-degree line, indicating perfect prediction accuracy, which would be very high. On the other hand, if MLR points were also plotted, they would show more scatter away from the 45-degree line.)Figure 1 scatter plot shows MLR models have low predictive accuracy. With RF models, the predicted COD values follow closely to the actual observed values, especially along the 45-degree line –which represents perfect prediction– with minimal deviation. This illustrates that the model accurately predicts river pollution levels.

CONCLUSION

This study illustrates the efficiency of machine learning methods, notably Artificial Neural Networks and Random Forests, in river pollution forecasting considering industrial discharge. The results demonstrate that sophisticated algorithms employing non-linear modeling outperform traditional linear methods drastically because the structure of the environmental data is non-linear. Such a high level of precision in predictive capabilities facilitates proactive management of the environment – allowing the design of early warning systems, advanced monitoring systems, and guiding measures to be taken in the control of pollution. Despite some of the challenges of data quality, model interpretability, and other issues concerning the reliability of information presented, the application of machine learning represents a powerful approach to managing ecological destruction, protecting water resources, and promoting responsible industrial development for the betterment of the environment.

REFERENCES

1. Tandon, B., & Thakur, M. (2025). An Overview of Adaptive Signal Processing Methods for 6G Wireless Communication Networks. International Academic Journal of Science and Engineering, 12(1), 12–15. https://doi.org/10.71086/IAJSE/V12I1/IAJSE1203
2. Hasan, M. S. (2024). The Application of Next-generation Sequencing in Pharmacogenomics Research. Clinical Journal for Medicine, Health and Pharmacy, 2(1), 9-18.
3. Anand, M. D. (2024). Design and Development of Advanced Mechanical Systems. Association Journal of Interdisciplinary Technics in Engineering Mechanics, 2(1), 1-6.
4. Usikalu, M., & Okafor, E. (2025). Strategic Innovation Models for Enhancing Organizational Agility in the Knowledge Economy. International Academic Journal of Innovative Research, 12(1), 8–13. https://doi.org/10.71086/IAJIR/V12I1/IAJIR1202
5. Castillo, M. F., & Al-Mansouri, A. (2025). Big Data Integration with Machine Learning Towards Public Health Records and Precision Medicine. Global Journal of Medical Terminology Research and Informatics, 2(1), 22-29.
6. Hawthorne, E., & Fontaine, I. (2024). An Analysis of the Relationship Between Education and Occupational Attainment. Progression Journal of Human Demography and Anthropology, 1(1), 22-27.
7. Shah, V., & Bansalm, T. (2023). Multidisciplinary Approaches to Climate Change Monitoring Using Cloud-based Environmental Data Systems. In Cloud-Driven Policy Systems (pp. 25-31). Periodic Series in Multidisciplinary Studies.