PREDICTIVE MODELING OF POLLUTION IN RIVER BASINS USING MACHINE LEARNING TECHNIQUES

HASLINDA BINTI ABDUL SAHAK

MCS241004_MCST 1043

CHAPTER 1

UNIVERSITI TEKNOLOGI MALAYSIA

NOVEMBER 2024

CHAPTER 1

INTRODUCTION

1.1 Introduction

Thus, pollution of river basins is among the many challenges faced by mankind in today's world, with some being very urgent like the effects they have on ecosystems and human health. Their wider socio-economic development is also affected in all the regions impacted. The fact is that now, anthropogenic and environmental changes have worsened the state of water pollution globally. The use of traditional monitoring methods like manual sampling and laboratory analysis is time-consuming as well as labor-intensive and has ranged limits in terms of space and duration.

With advanced machine-learning techniques, these technologies can learn to counter such challenges. The historical data along with the different meteorological parameters and relevant input factors assists the modeling of the air-pollution future prediction and gives remedial management strategies and pollution control measures for effective insights.

The machine learning technology adopts new speed in its evolving shape into a critical discovery tool in predicting water quality and the management of pollution in any river basin. Its ability to create models over complex datasets, recognize patterns,

and make predictions of monitored outcomes has made it a valuable tool for modeling programs that tend toward issues like water quality. Previous study is providing emphasis on how machine learning applications and deep learning applications can be useful for the prediction of water parameter reservoirs and events of associated pollution. For instance, Li et al. (2023) applied various machine learning techniques, namely, Random Forest, Support Vector Machines, and Gradient Boosting. These were quite effective in predicting some water quality parameters such as pH, dissolved oxygen, and turbidities. Zhang et al. (2022) predicted algal blooms in lakes through a deep learning model using Long Short-Term Memory (LSTM) networks.

Thus, this study aims to develop an effective machine learning model capable of predicting pollution levels in river basins within the country; it goes beyond technology improvements. Future trend predictions on the levels of pollution would help policymakers take prompt and directed actions on the conservation of these resources against pollution and even for public health safety.

1.2 Problem Background

River Basin Pollution, an important environmental problem, has severe effects on ecosystems, human health, and socio-economic development. Increasing anthropogenic activities, together with the effect of climate change, have compounded most parts of the world with problems of water pollution.

Industrial activities often pollute water bodies: in most cases, this is relevant to manufacturing, mining, and energy production activities. Metals, toxic chemicals, and organic pollutants are some of these pollutants that may enter water bodies. An instance of this is when the textile industry releases dyes and chemicals, whereas mining operations may account for the release of heavy metals to water bodies (Li et al., 2023).

Fertilizers and pesticides are part of non-point source pollution brought about by unsustainable agriculture practices. These get washed off from agricultural fields with water runoff into adjacent water bodies by rain, contaminating the water bodies and degrading the water quality. Further, toxic algal bloom and depletion of oxygen that arise as a result of greater nutrient loading from fertilizers may cause eutrophication (Zhang et al., 2022).

For rain that pours down in cities, it does not fall straight down to be absorbed in the ground as that of forests and fields. Instead, it flows over roadways, rooftops, and sidewalks. Such surfaces generate stormwater runoff, or water that cannot find its way through the soil. This runoff collects pollutants as it travels and eventually flows into storm drains and streams, where those pollutants are carried into the waterways.

Statistics seem to indicate that climatic change is defining changes in the pattern of rainfall. The occurrence of extreme events seemed to increase even with temperature. Such changes are worsening the scenario of water pollution by increasing the quantity and intensity of runoff, contributing to erosion and nutrient loading. High temperatures are also converted into faster speeds of chemical reactions in the water; thus, increasing pollutant concentrations.

Traditional methods of water quality monitoring, such as manual sampling and laboratory analysis, are very time-consuming, labor-intensive, and sometimes limited in spatial and temporal coverage. These do not succeed in capturing real-time variations in water quality and, therefore, contribute to delays in response to pollution events. The typical methods are often expensive, particularly for very large-scale monitoring programs.

1.3 Problem Statement

Pollution in river basins is a serious environmental issue, results in many ecosystem disasters, causes human health problems and thus hinders socio-economic

development. Traditional ways of monitoring water quality, e.g. manual sampling and laboratory analysis, often fail to be timely, accurate, and spatially comprehensive. Such shortcomings affect successful pollution management and the decision-making process.

Conventional techniques for monitoring are based on manual sampling and laboratory analysis, which are time-consuming and labor-intensive. Real-time variations in water quality may not be detected by these methods, resulting in slow reactions to pollution incidents. Furthermore, such methods may not be employed over extensive areas, leaving many of the spatial distribution aspects of pollution undiscovered.

Water pollution in river basins is a result of the interaction of various factors such as hydrological, meteorological, and anthropogenic. Anthropogenic impacts vary so involving interaction in a complex manner that predicting pollution levels becomes quite impractical. The successful development of predictive models requires the identification and quantification of all critical factors and their interactions.

It should be possible for such machine learning models to be integrated into the current water quality monitoring and management systems. Create simple, user-oriented tools and interfaces that will be easy for water resources managers and policymakers to use. Also, models must inter-operate with the current data systems and decision-making processes.

1.4 Research Questions

In this study, a strong machine learning model to predict pollution levels in Malaysian rivers will be developed relying on historical water quality data, meteorological factors and other parameters. The research questions are:

1. What are the effective skills for collecting and pre-processing high-quality, relevant, and consistent water quality data from the Department of Statistics Malaysia (DOSM) to train an accurate model and produce efficient predictions?

2. Which of the machine learning algorithms such as Random Forest or LSTM would work best in the prediction of pollution levels in Malaysian river basins, and how can these models be optimized for high accuracy and generalizability?

3. What way does the use of appropriate metrics (for example RMSE, MAE, R-squared) to evaluate developed models, and then deploy the same for real-time or near-real-time predictions for making effective decisions and intervention strategies?

1.5 Research Aim

The project intends to construct an advanced machine-learning model to predict pollution levels in the different river basins within Malaysia. This model would further help to predict future trends in water quality using historical trend data, meteorological data, and other variables that can be good for making rational pollution management and mitigation strategies. Thus, this study will bring sustainable management in the long-term Malaysia's water resources through proactive measures for ensuring water quality and in turn people's health.

1.6 Research Objective

The primary aim of this research is to establish a resilient machine-learning model for predicting the levels of pollution in Malaysian river basins for regulating timely and sound preventive measures against pollution to ensure water resource protection.

Objectives:

1. To collect historical water quality data with preprocessing before being made available at the Department of Statistics of Malaysia (DOSM) to ensure its quality and consistency.

2. To develop at least two machine learning models: Random Forest and LSTM, from pre-processed data, to achieve an initial accuracy greater than or equal to 75%.

3. To evaluate the created models according to appropriate dimensions (RMSE, MAE, R-squared) and optimize them for better accuracy.

1.7 Scope of Research

This research aims to build a machine-learning application for predicting pollution levels in river basins in Malaysia that takes on the following scope:

1. A data collection: Sources of historical information for water quality values, weather data, and other relevant parameter-related data include the Department of Statistics Malaysia and reputable others.

2. Data Preprocessing: Data cleaning: refers to the preprocessing of these somewhat diverse first data records missing values, outlier information, and inconsistencies.
3. Feature Engineering: Construction of appropriate features from the raw data like temporal features, hydrological features, and socio-economic indicators.
4. Model Building and Training: Varying machine learning algorithms and their application for training appropriate models on the willing, prepared dataset, these are Random Forests, Gradient Boosting, and LSTM.
5. Model Evaluation: Performing evaluations of the model that could be based on metrics such as RMSE, MAE, R squared, or through statistical tests.
6. Prediction and Visualization: Future forecasting from the trained model on pollution levels and the drawing of results.

1.8 Significance of Research

This research on Predictive Modeling of River Basin Pollution using Machine Learning Techniques is poised to revolutionize water resource management and environmental sustainability. By leveraging the power of machine learning, this study aims to develop advanced models that can accurately predict pollution levels in river basins.

Through accurate predictions, we can enhance water quality monitoring by early detection of pollution events, enabling timely interventions to mitigate environmental degradation. The insights gained from these models will empower policymakers and water resource managers to make informed decisions regarding water allocation, pollution control measures, and infrastructure investments.

Furthermore, by understanding future pollution trends, we can develop sustainable water resource management strategies that prioritize the health of our ecosystems and the well-being of communities. This research will also contribute to the advancement of machine learning applications, particularly in environmental sciences, opening up new avenues for innovation and problem-solving. Ultimately, the successful implementation of these predictive models can help safeguard public health, protect ecosystems, and ensure the sustainable management of our water resources.