PREDICTIVE MODELING OF POLLUTION IN RIVER BASINS USING MACHINE LEARNING TECHNIQUE

HASLINDA BINTI ABDUL SAHAK

MCS241004_MCST 1043

THESIS

UNIVERSITI TEKNOLOGI MALAYSIA

JANUARY 2025

ACKNOWLEDGEMENT

In preparing this project report, I was in contact with many people, researchers, academicians, and practitioners. They have contributed to my understanding and thoughts. In particular, I wish to express my sincere appreciation to Professor Madya Ts. Dr. Mohd Shahizan bin Othman for their guidance, advice, and motivation. Without their continued support and interest, this project report would not have been the same as presented here.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

The major reason for the pollution in river basins is that those are becoming dangerous threats to sustainability and health. This research attempts to develop a prediction model for forecasting pollution levels in Malaysia's river basins using efficient machine learning techniques. The model would use historical water quality data, weather data, and other relevant factors for future trends. It shall assist in providing insightful recommendations for efficient pollution management and mitigation measures. This project is intended to pave the way for sustainable management of water resources by trying to enable proactive approaches to protecting water quality and public health. The main objective of this research is to develop and evaluate highly advanced machine learning models in predicting water pollution within Malaysian rivers. Holistic dataset derived from historical water quality data include meteorological factors and some other relevant parameters acquired from DOSM. A critical part of the research work was the collection and preprocessing of water quality data from DOSM with a great deal of caution, considering accuracy, consistency, and suitability for machine learning model training. This has used some heavy data cleaning and preprocessing methodology for handling missing values, outliers, and inconsistency in the dataset. A range of machine learning algorithms was studied and assessed on their capabilities with respect to effective predictions of pollution levels. Among them were well-known powerful methods like Random Forest and Long Short-Term Memory-LSTM networks. It involved extensive tuning of hyperparameters to optimize each model's performance with the dual goal of maximizing its predictive accuracy and generalizability. Performance for these developed models is investigated rigorously with a suite of relevant evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. This was to ascertain the most appropriate and reliable model that could deliver real-time or near-real-time pollution level prediction with higher accuracy.

ABSTRAK

Punca utama pencemaran di lembangan sungai adalah kerana ia menjadi ancaman berbahaya kepada kemampanan dan Kesihatan. Penyelidikan ini bertujuan untuk membangunkan model ramalan bagi meramalkan tahap pencemaran di lembangan sungai Malaysia menggunakan teknik pembelajaran mesin yang cekap. Model ini akan menggunakan data kualiti air bersejarah, data cuaca, dan faktor relevan lain untuk meramalkan trend masa depan. Ia akan membantu dalam memberikan cadangan yang bernas untuk pengurusan pencemaran yang cekap dan langkah-langkah mitigasi. Projek ini bertujuan untuk membuka jalan bagi pengurusan sumber air yang lestari dengan berusaha untuk membolehkan pendekatan proaktif untuk melindungi kualiti air dan kesihatan awam. Objektif utama penyelidikan ini adalah untuk membangunkan dan menilai model pembelajaran mesin yang sangat maju dalam meramalkan pencemaran air di sungai Malaysia. Set data holistik yang diperoleh daripada data kualiti air bersejarah termasuk faktor meteorologi dan beberapa parameter relevan lain yang diperoleh daripada DOSM. Bahagian penting penyelidikan ini ialah pengumpulan dan prapemrosesan data kualiti air daripada DOSM dengan sangat berhati-hati, dengan mengambil kira ketepatan, konsistensi, dan kesesuaian untuk latihan model pembelajaran mesin. Ini telah menggunakan beberapa kaedah pembersihan data dan prapemrosesan yang berat untuk mengendalikan nilai yang hilang, outlier, dan ketidakkonsistenan dalam set data. Pelbagai algoritma pembelajaran mesin telah dikaji dan dinilai berdasarkan keupayaan mereka untuk meramalkan tahap pencemaran secara berkesan. Antaranya ialah kaedah yang berkuasa dan terkenal seperti Hutan Rawak dan Rangkaian Memori Jangka Pendek-LSTM. Ia melibatkan penalaan hiperparameter yang meluas untuk mengoptimumkan prestasi setiap model dengan matlamat ganda untuk memaksimumkan ketepatan ramalan dan kebolehpercayaan. Prestasi bagi model yang dibangunkan ini disiasat dengan ketat menggunakan set metrik penilaian yang relevan: Ralat Kuasa Dua Min (RMSE), Ralat Mutlak Min (MAE), dan R-kuasa dua. Tujuannya adalah untuk menentukan model yang paling sesuai dan boleh dipercayai yang boleh memberikan ramalan tahap pencemaran masa nyata atau hampir masa nyata dengan ketepatan yang lebih tinggi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BOD	-	Biochemical Oxygen Demand
COD	-	Chemical Oxygen Demand
CWQM	-	Continuous Water Quality Monitoring
DOE	-	Department of Environment
GANs	-	Generative Adversarial Networks
IoT	-	Internet of Things
ML	-	Machine Learning
NH-N	-	Ammoniacal Nitrogen
PCA	-	Principal Component Analysis
TSS	-	Total Suspended Solids
MWQM	-	Manual Water Quality Monitoring
WQI	-	Water Quality Index

LIST OF SYMBOLS

R^2	-	R-squared

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 labour-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.

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.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Predictive modeling of river basin pollution in which machine learning techniques are used has attracted a lot of attention recently due to their capability to provide actionable insights for conservation and policy-making in the environment. The chapter reviews existing literature to build a foundational understanding of methodologies and frameworks in similar studies. Thus, the review includes data collection, pre-processing, machine learning techniques, prediction and modeling, insights and applications, and alibis with future directions. This chapter also provides information relating to water quality and the classification of river in Malaysia. Three (3) existing methods have been identified for water quality monitoring which are Manual Water Quality Monitoring (MWQM), Continuous Water Quality Monitoring (CWQM) and IoT real-time water quality monitoring.

2.2 Data Collection

Diverse and high-quality datasets are needed for effective predictive modeling, an indispensable process for the modeling of pollution in river basins. The most vital phase of data collection is quality data collection, which maximizes the reliability of machine learning models. This phase integrates different data types and enables a thorough understanding of pollution sources as well as their interactions. As mentioned above, there are three existing methods used in Malaysia that focus on water quality monitoring. There are over 189 river basin systems from 1800 rivers, of which the Department of Environment (DOE) Malaysia has maintained over 146 basins. Six parameters have been chosen to calculate the water quality index (WQI) and used to classify the river in Malaysia. Those parameters are pH, Biochemical Oxygen Demand

(BOD), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Ammoniacal Nitrogen (AN) and Suspended Solid (SS).

2.2.1 Water Quality Monitoring

Water quality monitoring involves measuring several physical, chemical, and biological attributes to judge the health of a river basin. Of these, the pH plays a very important role as an indicator for such types of measurements in assessing the quality of water for aquatic life as well as for human beings. Any change in pH, whether becoming highly acidic or highly basic, is largely damaging to organisms and very much alters chemical reactions in it. Another important thing is dissolved oxygen (DO), as it protects the life of aquatic organisms. The decrease in values of DO is a direct result of organic pollution, as it will show the quality deterioration of a water body. Conversely, Biochemical oxygen demand (BOD) measures the volume of oxygen that microorganisms require to break down organic matter and thus serves as an indirect indicator of the level of organic pollution from exogenous factors.

Chemical oxygen demand (COD) further examines the overall amount of oxygen necessary to oxidize any organic and inorganic compounds, hence it extends the study to render a more complete meaning of pollution form. Total suspended solids (TSS), give one idea of the water clarity and sedimentation rates that occur; assessment of heavy metals such as lead, mercury, and arsenic depend on their toxicity and long-term damage to the environment (Chapra, 2008). These parameters are monitored in time series for a dynamic view of the water quality trend required in modeling. Regular monitoring and multi-point sampling are stressed by Khan et al. (2020) and Zhang et al. (2019) as mandatory for data to be representative across river basins.

2.2.2 Sources of Pollution

Identifying pollution sources, be it agricultural runoff, industrial discharges or urban wastes, is an important aspect. As essential is the identification of pollution sources because it helps the researchers in knowing where contaminants come from

and how they interact in the watershed. Considerable sources of industrial effluents include a myriad of pollutants, which may be chemicals as well as oils and heavy metals. As noted by Smith et al. (2018), point and non-point pollution sources are also determining factors of water quality dynamics; thus, effort must be made towards comprehensive data collection.

It has also been revealed that agricultural runoff brings additional fertilizer, pesticides, and organic matter to the river through surface runoff effects during rainfall. Urban wastewater, normally untreated within the developing areas, may also contribute nutrients, pathogens, and other pollutants to the river environments. Mining activities have a direct impact on physical and chemical pollution, like sedimentation and heavy metals leaching, in the environment. Finally, other natural sources such as soil erosion, decaying vegetation, and natural deposits of minerals also feature in the contributions to surface water quality. Triage of these sources is important in identifying the major sources of pollution and intervention measures to be taken (Novotny, 2003).

2.2.3 Geographic and Hydrological Data

This third pillar of data collection includes geographic and hydrological information, which together even more provide a spatial and environmental context to the actual water quality and pollution source data collected. This information is indeed geospatial: it generally refers to such features as river topography, land use pattern, and climatic condition, all essential in predictive modeling. For example, steep-sloped terrain increases runoff velocity, while flat terrains allow for larger sediment accumulation in a river. The other critical factor is flow in such rivers, determining how greatly a river can bank pollutants. High flows will usually dilute pollutants and disperse them over large areas while low flows lead to the accumulation of pollutants in a given location. The patterns of rainfall also play a significant role in driving these polluted waters since they determine the volume and intensity of runoff that will flow into the river.

Data on land use and cover population gives a meaningful relationship between human activities whereby connection comes through urbanization, agricultural intensification, and deforestation vis-à-vis water quality changes. Such as nutrient loading by agricultural land and organic and chemical pollutants from urban location wastewaters (Ward and Robinson, 2000). The investigations such as that of Singh et al. (2021) illustrate this better when marveled by the fact that GIS integrates with learning machines for improved spatial analysis.

With the systematic collection and integration of these diverse datasets, researchers have created a widely comprehensive knowledge about the pollution dynamics in river basins. A base for machine learning, the multi-faceted data is used to detect complicated patterns, predict, and inform future pollution trends, and develop sustainable management practices for river basins.

2.3 Data Preprocessing

Data preprocessing constitutes the critical component of the machine learning pipeline that attains high-quality and valuable data for analysis and modeling. It refers to a set of techniques and processes, worthy of purification or cleaning raw data to be processed further for better performance and accuracy of the machine learning models. It cannot be stressed enough that quality data will always prove directly effective in rendering predictive models fail or succeed. In the context of any river basin pollution monitoring study, data preprocessing becomes a requisite factor. Often, environmental data are compiled from various sources, thus, in one way or another, they may have been subjected to inconsistent representation, missing entries, and outliers that finally tend to cause the model to fail to produce expected predictive outputs in the future.

Using appropriate cleaning methods like dealing with missing values and filtering out outlier data, researchers can ensure the robustness and reliability of their data. Furthermore, feature engineering was the new creation of features, and relevant importance is critical in uncovering the data's hidden patterns and relations. Data categorization is essentially the last part of preprocessing. It involves the actual splitting of the dataset into model training, validation, and test samples. Thus,

preprocessing turns out to be a crucial step in preparation for such learning-based models in river basin pollution monitoring. Transformation into a proper and clean data source makes the modeling accurate and therefore the results financially viable for environmental management and sustainable development.

2.3.1 Data Cleaning

Data cleaning is a foundational step in the process of predictive modeling, which will guarantee the integrity of results before it is used in analysis or for entry into machine learning models. This step involves identifying and correcting inconsistencies, mistakes, or the absence of data within that dataset to facilitate a strong foundation for the proceedings (Han et al., 2011).

Missing values in a data set significantly threaten the accuracy and performance of the machine learning model itself. Therefore, handling such missing values is a pre-condition in any data cleaning step. To take care of this problem, there are several techniques, mainly imputative methods, which tend to calculate and replace the missing values from other data points. In mean imputation, the mean of a feature is taken as the value of all missing values of that feature. Median imputation takes into account the median value of a feature, and mode imputation takes the most common value as an imputation value. More advanced procedures, for example, k-nearest neighbors (KNN) imputation consider the values of k-nearest data points in the feature space to impute the missing values by referring them to other values (Hastie et al., 2009). MICE is a multiple imputation approach, which creates multiple imputed datasets, analyzes each separately, and concludes a pooled estimate for increased accuracy. The selection of a method is dependent on the dataset characteristics that describe the distributions of the missing values.

The next step of data cleaning is outlier removal, with the intention that outliers are the data points that differ considerably in general trends. Outliers can be deletable from statistical analyses and can make some machine learning models perform disproportionately poorer than other models. Such detection and treatment are usually done statistically using the Z-score method or using the interquartile range (IQR). The

Z-score method registers outliers based on how far away they are from the mean in standard deviation units with data points superior to some critical value (say 3) considered outliers. The IQR, instead, considers the distribution of the central 50% of the data. Outliers are defined as values that are either below the lower quartile minus 1.5 times IQR or above the upper quartile plus 1.5 times IQR (Aggarwal, 2015). Outliers may be omitted from the dataset or replaced with other values that seem more appropriate or might decrease the effect of their values on the analysis.

Normalization and standardization can be considered as additional techniques in data cleaning, which scales numerical features on a common range or distribution. For instance, normalization scales data between a minimum and a maximum value, for example, between 0 and 1, which means that all features would contribute per unit weight to the model. An example of this is min-max normalization. Standardization, on the other hand, means changing the data's mean to 0 and its standard deviation to 1, usually by Z-score normalization. Both of these operations prevent certain features with larger ranges or units from overpowering such learning mechanisms. In this sense, the improvement in performance or stability of the model increases (Han et al., 2011). Such scaling of data is more critical for distance measurement-based algorithms, for instance, the k-nearest neighbors and support vector machines.

Addressing missing values, outliers, and scaling aspects in these conditions means data cleaning will work to ensure that the data can be moved toward accuracy, consistency, and an effective predictive model. This is an important first step in generating accurate and sensible predictions of pollution in river basins.

2.3.2 Feature Engineering

Feature engineering is quite pragmatic in developing any predictive models, which essentially means creating new features or transforming existing ones to enhance machine learning model performance. It includes discovering relevant important variables, identifying temporal patterns and characteristics related to geographies, and maximizing the relationships for which the model is intended. It involves the definition of significant variables, depending on the temporal patterns,

and geographical traits to maximize what the model understands of all the relationships for which it is designed. Feature Engineering is a never-ending process through the lifetime of models.

Feature engineering is the initial step where selection of the important features is made considering selection of most relevant features affecting pollution levels in river basins. It combines knowledge from domain, statistical analyses and a number of techniques from machine learning. Examples of such features include industrial discharge, agricultural runoff, or urban wastewater for which knowledge from domain could inform experts in identifying those features of theoretical or practical association with pollution. For statistical methods, such as correlation analysis, determine the strength and direction of the relation between features and the target variable, providing quantitative evidence used in selecting features. Machine learning models such as random forests or even gradient boosting provide feature importance scores that rank variables for their predictive power. Otherwise, together these approaches warrant the inclusion of the most meaningful variables in the data set, thus improving the accuracy as well as interpretability of the model (Guyon and Elisseeff, 2003).

Temporal feature extraction is most relevant to river basin pollution data having time variations. Time-related features, for example, conductive trends, seasonal variations, or periodic behaviors, which carry important dynamics for understanding pollution, can be very significant when it comes to time-dependent aspects like river basin pollution data. Examples will be the time of year, month, or even season, all of which can indicate seasonal differences in rainfall, agricultural activities, or industrial discharges. It can be through Fourier analysis or time series decomposition that extracted seasonal patterns can be observed on time-stamped data. The trend-analysis methods mostly used are moving averages or exponential smoothing, useful to trace long-term trends in pollution levels over time. Given how temporal features work, such models can thus learn how pollution variables vary at a certain location and time by cyclical changes at certain periods in environmental factors (Chatfield, 2003).

Such integration of geospatial feature makes the feature engineering process more complicated and deep as spatial attributes come into play very much when predicting pollution levels. River basins have that strong spatial parameter determining

the location, topography, and land use. Geospatial feature engineering is the step of integrating in spatial data, indeed as the distance from the source of pollution or even their position only at their monitoring stations, in order to capture location effects on dispersion of pollution. I can assess distance between sources and monitoring points with regard to the customized demand of understanding how proximity causes pollution effects. Distance matching sources and monitoring points can assess how proximity in pollution levels might be effective by the calculated distances. Further extension by adding altitude data gives it level heading enter and take topography effects, which have differences in water flow and carried pollutants under two zones such as higher elevation and lower elevations. Other connected land uses could include areas that are urban, agricultural, and forested, for instance, in quantifying possible impacts caused by human activities or natural landscapes on water quality. Bringing in these geospatial attributes allows the model to better understand spatial and environmental predisposition to pollution (Goodchild et al., 1992).

Feature engineering through the discovery of important variables, temporal profile extraction, and geospatial feature infusion will augment the dataset and improve the power of machine learning models. It uses domain knowledge, temporal patterns, and spatial dimension to ensure the model captures the multiple ways pollution dynamics in river basins and the enhanced predictive capabilities for accurate and detail-rich results.

2.3.3 Data Splitting

Data splitting, which is the most important part of data preprocessing for machine learning projects, provides effective model training, validation, and testing for genuine results that can be generalized. In this context, as per predicting pollution levels in river basin systems, the process involves dividing the dataset into three sections: training, testing, and validation datasets (Hastie et al. 2009).

The model development entirely relies on training data. A training dataset can be defined as the subset of data that helps a machine learning model learn the patterns, relationships, and trends in a given dataset. At this point, the model fits its parameters

so that errors can be minimized, leading to improved predictive ability, based on incoming data. The size and quality of the training dataset are critical to ensuring the model's ability to generalize to unseen data (Géron, 2019).

The testing dataset is employed to evaluate the model's performance on data that it has not encountered during training. By assessing the model's accuracy on this unseen dataset, researchers can estimate how well the model is likely to perform in real-world scenarios. This evaluation provides an unbiased measure of the model's predictive capabilities and helps identify any issues related to overfitting or underfitting. The testing dataset should be representative of the overall data distribution to yield meaningful performance metrics (Han et al., 2011).

The dataset validation is vital in model tuning and optimization for hyperparameters; however, hyperparameters control the model's learning such as learning rates, how many layers the neural networks have, or how deep the decision tree can be. The validation set is against which all model configurations are evaluated and the one that does best in the validation set is selected to prevent overfitting, whereby overfitting occurs when models become too specific to training data and therefore do not test well on new data. Among those model configurations, it finds which one scored best with the validation dataset bringing researchers one step closer to finding the right balance between model complexity and generalizability (Kuhn & Johnson, 2013).

Datasets in machine learning are trained, tested, and validated. The training dataset helps the model learn emerging patterns, the testing dataset evaluates its predictive performance, and the validation dataset fine-tunes the model to prevent overfitting and, as a result, optimize performance. It ensures the split of data for developing robust and reliable predictive models on the level of pollution in river basins.

2.4 Machine Learning Techniques

Machine learning is about designing as well as writing algorithms for computers to learn directly from data and hence predict or decide without being specifically programmed. One of the examples where machine learning plays a very important role is in predicting pollution in river basins, where it could analyze complex patterns and trends within the data that would not otherwise be capable of analysis using traditional approaches. Machine learning techniques can be classified into three categories; Supervised Learning, Unsupervised Learning, and Time-Series Analysis. All these broad categories further classify into particular techniques or models into which they are incorporated.

2.4.1 Supervised Learning

Supervised learning is a type of learning using labeled data, unlike unsupervised learning in which the input data (features) is not associated with its output, here the input data is associated with its output labels (target variables). The model can learn and map the input features to the output variable, enabling it to predict output within new, unseen data. Regression Models serve to predict continuous numerical values, such as the pollution level in a river basin. Various regression models such as linear regression and polynomial regression can be drawn depending on the degree of complexity of the relationship between the features and the target variable. Linear regression assumes that all the input features have a linear relationship with the output, whereas polynomial regression can capture more complex nonlinear relationships (James et al., 2013).

Decision Trees are considered those tree-like models which take decisions based on a set of rules derived from the input features. They are useful in both classification and regression tasks. In the case of pollution prediction, decision trees can help identify the most critical factors causing pollution levels since they split the data based on those factors (Quinlan, 1986).

Random forests are an ensemble of learning techniques based on a union of decision trees, whose accuracy and overfitting are intended to be reduced. Much of their power comes from aggregating many predictions for building a more robust model that is less sensitive to the random variance for individual trees. Random forests are most effective when there are complicated environmental data sets characterized by hundreds, thousands, or more than a thousand interacting variables (Breiman, 2001).

Gradient Boosting Machines (GBMs) are another ensemble learning method that builds models in succession with the models adding to each other and focusing on correcting errors made in the previous models. It is strong because it serves as a combination of several weak learners to form a strong prediction model. It is of great use in cases that need a lot of predicting accuracy; it has been successfully applied in different environmental modeling tasks (Friedman, 2001).

2.4.2 Unsupervised Learning

Unsupervised learning algorithms are built for analyzing datasets without labelled output variables. It is not creating categories or outcomes; these algorithms will rather try to discover concealed patterns, structures, and associations in the data. So, it becomes useful for exploratory data analysis. Unsupervised learning can bring to light latent tendencies within river basin pollution prediction and further cluster similar points based on specific characteristics (Bishop, 2006).

Unsupervised learning is indeed used widely to perform this operation that is termed clustering-a collection of similar data points against feature dimensions that they share. Clustering techniques can be K-Means as well as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) which help reveal the group of similar pollutant patterns. Either define several clusters and optimize minimization of variance within these groups as with K-Means, or target a dataset that is noise and density varying in clusters like DBSCAN, which best suits an analysis of relevant pollution data despite the irregularities or outliers included in it (Hastie et al. 2009).

Another critical aspect of unsupervised learning is dimensionality reduction, which aims to reduce the number of features of a dataset while maintaining the most valuable information. Dimensionality reduction becomes especially important when working with high-dimensional data since it makes the analysis easier and speeds it up computationally. Principal Component Analysis (PCA) is a well-known technique in dimensionality reduction that transforms the data into fewer uncorrelated components and achieves maximum variance. Like t-Distributed Stochastic Neighbour Embedding (t-SNE), this method is nonlinear and does a better job of visualizing high-dimensional data by preserving the local relationships in a lower-dimensional space (Maaten and Hinton, 2008).

Researchers are able to acquire a deeper understanding of the structure and attributes of emissions data through application of clustering and dimension reduction techniques. These methodologies serve not only for the understanding of complexities in datasets but also serve the importance of further modeling and decision-making in environment management.

2.4.3 Time-series Analysis

Time-series analysis refers to techniques being utilized in the analysis of data collection, over time and space, such as a few daily, weekly, or monthly recordings of river basin pollution. This technique is for recognizing the patterns, trends, and season influences on this data to allow better predictive models and better decisions also (Chatfield, 2004).

ARIMA, which stands for autoregressive integrated moving average, is the most commonly used model in time-series analysis. The three parts of ARIMA are autoregression (AR), integration (I), and moving average (MA). The AR component models the relationship between an observation and its several previous values; the I component accounts for non-stationarity by differencing the data; and the MA component captures the dependency between observation and residual errors from previous predictions. This model has been widely used to forecast pollution levels, as it can sufficiently capture linear trends as well as seasonal variations (Box et al., 2015).

As a result, the Long-Short Term Memory, an advanced form of recurrent neural network, has always been used in the case of more complex and nonlinear time-series data. LSTMs are capable of long-term dependencies and temporal dynamics, which are helpful in analyzing time series. Memory cells keep their states over time in these networks to build very complex relationships as needed for the prediction of pollution levels caused by many dependent factors over very long periods (Hochreiter and Schmidhuber, 1997).

Seasonal trend models which are another important class of time-series analysis models are concerned with a recurrent variation of recurring increase or decrease noise in pollution levels during the year for longer-term trend identification. Seasonal trend models include a technique to decompose observed time-series data into its components, which include seasonal effect, trend, and noise, thereby elucidating a clearer understanding of the underlying dynamics in the data collected (for example, Hyndman and Athanasopoulos 2018). Application of these methods could lead to a better understanding of the temporal patterns of pollution levels in river basins and would facilitate precise predictions supported by sound environmental management strategies.

2.5 Prediction and Modeling

Comprehensive prediction and modeling systems are thus fundamental in projecting pollution levels within rivers basins and assist researchers and policy makers in understanding pollution dynamics along with the design for effective treatment strategies. For instance, predictive models study the trends of historic and present data, patterns, and relationships, making it possible to predict future pollution levels accurately (Hastie et al., 2009). Simple statistical techniques, such as linear regression, to advance machine learning methods, including Random Forests and Deep Neural Networks, are major methods that such models employ in making sense out of complex environmental data. Random Forest model results, for example, put emphasis on the need for consideration of various environmental variables while deep learning frameworks, such as Convolutional Neural Networks and Long Short-Term Memory models, especially encapsulate spatio-temporal patterns in pollution data (Le et al., 2015).

Predictive modelling has come up with metrics using which one can wise enough evaluate the models in terms of how much accurate their prediction results are: to name a few Mean Absolute Error, Root Mean Square, and R-squared (Kuhn and Johnson, 2013). A process of cross-validation is used to make sure that actual performance of the model has been assessed on unseen data, which minimizes its overfitting hazards and attains a measure of generalizability in the future. These would surely enhance prediction but would also tell us all about the causes and drivers of pollution, thus empowering management efforts to be data driven. A predictive model with strong computational methods and domain knowledge is a very viable approach to solving pollution-related problems in river basins as well as ensuring sustainable management of freshwater resources (Montgomery et al., 2015).

2.5.1 Short-term Predictions

Pollution levels can be predicted over short-term durations like days, weeks, or months, judgments that provide a basis for much-needed actionable insights into real-time monitoring and management of water quality. Such predictions can identify sudden shifts in water quality and intervene against possible pollution events. This type of analysis of water quality short-term trends informs stakeholders on the immediate application of mitigation measures against adverse environmental impacts, ensuring the safety and sustainability of water resources (Hastie et al., 2009).

The primary application of short-term predictions is the real-time monitoring of water quality indicators. Predictive models routinely read and process data capture of the present and past measurement of such indicators, they analyze and detect deviations or sudden shifts in pollution levels and trigger alerts for intervention. Another very important dimension of short-term or short-range prediction is capturing this seasonal variation from at least an annual pattern due to exogenous factors, such as rainfall patterns, temperature changes, and agricultural activities. With these seasonal trends, proactive plans and resource allocations can also be possible so that water quality will remain within acceptable limits even during critical periods (Hyndman and Athanasopoulos, 2018).

2.5.2 Long-term Predictions

Long-term predictions are forecasted for pollution levels for years to decades. These types of forecasts are essential in understanding the effects of different environmental and anthropogenic factors on water quality over time. These predictions will be able to provide ways into the future, leading to sustainable management strategies for water resources to remain resilient in times of change (Hastie et al., 2009).

A vital use of long-term forecasts would be to assess the effects of changes in land use-modified water quality due to urbanization, agricultural extension, or deforestation. Predictive models can illustrate the extent to which pollution sensitivity levels are determined by such environmental changes and therefore provide information on land controls that could mediate negative impacts. Long-term forecasts would also be used to gauge the effectiveness of policy interventions, such as pollution control measures and water conservation programs, which would allow flexibility and adaptation of policy strategies over time. At the same time, long-term predictions are significant concerning potential impact assessment for climate change on water quality changes for more or less precipitation, temperature, or varying maximum extreme weather events. They are requirements for adaptive management plans which are optimized to address problems that will emerge in the future guaranteeing sustainability in water resources (Hyndman and Athanasopoulos, 2018).

2.5.3 Model Validation

Model validation allows for modeling and prediction-making ensuring a reliable performance of machine learning models while forecasting levels of pollution. This simulated scenario can facilitate an evaluation of how well the model generalizes its predictions to novel or unseen data. This procedure includes testing the model with different data partitioning or through different trials to ensure that it does not overfit the training set of data but performs effectively in real-world situations. Effective model validation enhances the credibility and utility of predictions, making them

actionable in environmental management and decision-making: Kuhn and Johnson (2013).

To obtain valid soundness, cross-validation is typically adopted. The procedure is characterized by dividing the data into predetermined training and testing subsets, followed by various assessments at different points of the split. Apart from these, performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) are computed to quantify actual predictions and the reliability of the model. These measures may be important to know the relative strengths of the model, while they show the weaknesses in certain areas that can be improved for better prediction accuracy (Hyndman and Athanasopoulos, 2018).

2.5.3.1 Cross-validation

In machine learning, cross-validation is a common method used for evaluating the model based on its generalization capability to unseen data. It is almost always done by partitioning the dataset into training and testing sets, thus enabling a sound method of checking the accuracy and reliability of the model. Common approaches include k-fold cross-validation, which divides the dataset into k equal parts, training the model on k-1 folds and testing it on the remaining fold in an iterative manner, and leave-one-out cross-validation, which evaluates the model by training on all data points except one and repeating this process for each point. These methods are also very effective in overcoming overfitting and improving model generalization, assuring more effective real-world performance (Zhang et al., 2021b).

Cross-validation is effective for predictive modeling outcomes. Zhang et al. (2021b) found that cross-validation techniques do not just reduce overfitting, but help in identifying the best configuration of models by providing information about the performance of the model on different subsets of the data. Such structured evaluation gives sufficient information to test the model completely; hence one can be confident that the model's predictions are as well applicable in practice.

2.5.3.2 Error Metrics

Error metrics are important for evaluating the efficiency and precision of machine learning models. They are a scientific measure of how well the predictions by the model agree with the observed quantities, thus giving it credibility and relevance to the measure. Error Metrics most commonly used are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). The first one considers the average magnitude of errors in a set of predictions and indicates how close the prediction is to that of a human interpretation as to the model's accuracy. If, however, very large errors were to belong to the distribution of errors, then RMSE would typically be the metric that would be most relevant because of the much stronger weight given by it to larger errors. On the other hand, R-squared evaluates the proportion of variance in the dependent variable that is predictable from the independent variables, thus providing a more complete view of the model's explanatory power (Patel et al., 2020).

Patel et al. (2020) presented in-depth comparisons of the error metrics and cited a lot of their strengths and conditions of applicability in different situations. Their work notably states that MAE is appropriate in cases where interpretability is required while RMSE is warranted for datasets that present quite a large variability in errors. Using the R-squared metric is also for general goodness of fit testing in regression models. Hence, accurate error metrics can help practitioners' complete evaluations of machine learning models to yield more effective and reliable predictions.

2.6 Insights and Applications

This section gives detailed information on strategies and ways to manage environmental pollution. Machine learning provides an excellent tool for analyzing huge and complex datasets such as water quality indices and anthropogenic activities, as well as meteorological data, which are often nonlinear and interrelated. These researchers might be able to realize patterns and relationships among the variables that impact the level of pollution with a deeper understanding of water quality degradation

(Huang et al., 2020). Therefore, predictive models can simulate variations in pollution in terms of time and space for more proactive planning and management measures. These alternative scenarios could also include land use changes, industrial discharges, or climate conditions and would be an effective way of appraising the future impacts of different interventions and policies (Zhang et al., 2021).

Applications of this research extensive and impactful. These models can help policymakers and environmental managers frame and enforce evidence-based regulations capable of controlling pollution in river basins. Aside from identifying the exact place and time when pollution occurs, these models can help guide the cost-effective and efficient allocation of resources for the monitoring and mitigation of pollution (Gholami et al., 2021). With integration from real-time monitoring systems, it can even help build a full-fledged early warning system in which stakeholders can respond early to new pollution threats and mitigate the likelihood of effects on the ecosystem and public health (Yang et al., 2018). The insights yielded from these models may also help in reaching communities and stakeholders for education, thus raising awareness of their participation in water quality management. Predictive modeling hence serves not only in fitting settings for immediate pollution control but rather finishes the full cycle that leads to sustainability as development activities are well coupled with environmental preservation goals (Shao et al., 2019).

2.6.1 Pollution Hotspot Identification

Machine learning has important applications in determining pollution hot spots that are areas with higher incidence of contamination or more susceptible to pollution events. Application of predictive models from large amounts of data on water quality, land use, conditions of the weather, and even human activity can be used to identify high-risk areas within river basins or watersheds. Environmental agencies can then use the results to direct attention and reopening activities to the most vulnerable areas to pollution. For instance, Gupta et al. (2020) utilized k-means clustering algorithms to map pollution hot spots within the Ganges River Basin, which were earmarked as such to receive immediate attention. These findings are very important for concentrating

efforts towards the most critical areas, thus enhancing the effectiveness of pollution control measures (Gupta et al., 2020).

Predictive models are used in the first place to locate most effectively the high-risk zones along a river basin with respect to pollution levels, or places where pollution events are likely to occur. These models can forecast pollution levels in different regions based on combinations of factors such as industrial activities, agricultural runoff, and hydrological conditions. The capability helps create monitoring strategies for different water quality measures with time-scale remediation efforts. Earlier identification of zones is necessary to adopt proactive measures for pollution management to prevent possible irreversible damage to the environment (Kumar and Singh, 2020).

Seasonal variations in pollution levels have occurred where predictive models provide input regarding the other significant case. Pollution levels keep changing according to the seasons owing to rainfall or agricultural practices or change in water flow. For example, during monsoon seasons, heavy flooding runoff is increased from farms leading to the nutrient and pesticide pollution levels within the nearby water bodies.

Predictive models are capable of identifying these seasonal profiles, thereby allowing the authorities to target interventions at times of peak pollution. Understanding the temporal variation provides a more efficient allocation of resources, as an allocation may then coincide with periods of risk, for instance, at post-harvest times or heavy rainfall (Agarwal and Singh, 2018).

2.6.2 Policy Formulation

The insights that we got from predictive by machine learning models can highly contribute to the development of evidence-based policies and regulations on better environmental management. A model simply predicts future pollution trends by examining historical data; thus, it helps policymakers make regulations that are proactive instead of reactive. Jones and Taylor (2019) shed light on how machine

learning models have had a significant contribution toward the development of sustainable water management policies in the European Union. These models give documentary evidence that helps policymakers to make informed decisions in the design of policies that will reduce pollution while promoting sustainability in resource utilization.

Machine learning predicts and have capacities to develop data driven regulations for the reducing pollution into river basins. The models identify the different pollution sources and future trends prediction so that the regulations would be more specific and effective. For example, this type of analysis provides clearer views on pollution dynamics under different environmental conditions and could help in establishing the regulatory limits that are closer to reality. It is formulating policy-driving data in regulatory bodies, which would grant more targeted, efficient, and adaptable environmental standards to emerging pollution concerns (Kumar and Singh, 2020).

The use of predictive models in policy making is valuable when it comes to developing an early warning system for pollution events. A machine learning model, trained on environmental variables through emission, can predict future occurrences of pollution events, for instance, chemical spills, industrial discharge, or agricultural runoff beyond accepted levels. Such early warning systems allow authorities to take preventive or emergency measures when an event is nil rather than do so after it has happened, thereby improving chances of minimizing environmental damage. Early warnings also permit remediation measures, for instance, water treatment or containment, or public advisories, resulting in more effective timely interventions (Agarwal and Singh, 2018).

Predictive models facilitate targeted remedial actions by spotlighting region-specific pollution problems. Such models can help the policymakers prioritize resource allocation to areas that source pollution, the level of contamination, and effectiveness in already enforcing mitigation measures. They may even predict where those resources should be directed for targeted interventions that are more effective or severely affected by pollution, and would probably make a more significant impact on improving water quality. Such knowledge can shape streamlining remediation

processes and therefore increasing the quality of overall environmental governance. (Li et al., 2019).

2.6.3 Environmental Impact Assessment

Environmental impacts assessments (EIAs) are fundamental for measuring the effects that such predominantly industrial but sometimes agricultural and infrastructural, would have on the environment and its sensitive ecosystems, such as river basins. Predictive models integrated into EIA systems provide data-based forecasts on the expected environmental effects of projects, such as their eventual influences on water quality, biodiversity, and overall ecosystem health. For example, the work of Ahmad et al. (2021) demonstrates the potential of predictive analytics and their integration into EIA frameworks that allow for more precise and proactive impact assessments. These models display possible scenarios of the project, ranging from a broader time horizon of several decades to just a few years, and provide better informed decisions concerning the potential long-term ecological consequences of these scenarios.

Predictive modelling tools can help in modeling and assessing the different ecosystem health indicators that are affected by pollution. These model the changes in water quality, biodiversity, and other health conditions affecting aquatic life from pollution. All types of pollutants can be modeled, be they heavy metals, other nutrients, or both as they accumulate in a water body and determine how this will have an effect on fish populations or absorb plant growth. With this impact analysis, policymakers and environmental managers identify species and ecosystems reportedly at risk and begin developing appropriate mitigation strategies for reducing damage. Such an approach looks after the stability of an ecosystem while giving allowances for developing projects that may not want to harm natural habitats irretrievably (Li et al., 2019).

Predictive modelling of river basin management shall lead to improved sustainable strategies with human and environment needs harmonization. Furthermore, these models show how land-use changes, pollution, and climate factors

would impact the river basin over time. Moreover, by forecasting such effects from different management, it will be open for consideration as to which approaches to use with respect to water resources management, local community support, and biodiversity. For example, predictive models may be used to develop effective flood control facilities and pollution management measures and long-term water conservation efforts for sustainable river basins (Kumar and Singh, 2020).

The predictive models are further used to assess the socioeconomic impacts of pollution. The economic effects can be predicted when such things as water contaminants and environmental degradation would be related to impacts on human health, agriculture, tourism, and other sectors of the economy. These would include estimating the economic costs due to waterborne diseases, decreases in crop yields, or a decline in tourism due to polluted sources of water. By understanding these effects, development practitioners can formulate policies that protect communities against the negative socioeconomic impacts of pollution while minimizing environmental damage's economic costs (Agarwal and Singh, 2018).

2.7 Challenges and Future Scope

There are numerous opportunities for better environmental resource management through the application of machine learning (ML) techniques for predicting pollution in river basins. However, in all these opportunities, there are problems that need to be addressed. One of the first and leading causes of such challenges is the complexity and variability of environmental systems. River basins are dynamic systems, living with a number of factors such as land uses, climate changes, pollution sources, hydrological conditions, and the like. These cause very complex interactions, making it very difficult to develop accurate and robust models.

In addition, another significant obstacle is the availability and quality of data. Most machine learning models require large, quality datasets, which may not be readily available when a region is less monitored in terms of infrastructure (Li et al., 2019). A second challenge is the model interpretability. Advanced techniques such as deep learning are usually very accurate, but they operate as 'black boxes,' as the

relationships of input variables to the output are unclear and would thus affect regulatory and policy-making decisions (Ahmed et al., 2020).

There is great scope for growth in this field despite such difficulties. Future studies could perhaps have a better look at augmenting data acquisition; for instance, improved remote sensing techniques or sensor deployment in hitherto under-monitored areas might be areas of future work. Further, improved integration of multi-source data such as satellite imagery, socio-economic factors and real-time monitoring systems may significantly improve the accuracy and applicability of models predictive to this area (Kumar and Singh, 2020). Advancements in explainable AI (XAI) could be another way to enhance the transparency and trustworthiness of these models for decision makers. Future research, however, could look into the application of ensemble models that combine many machine learning algorithms to solve different types of data or prediction problems and thus come up with better-comprehensive and reliable results (Agarwal and Singh, 2018).

2.7.1 Data Challenges

Data challenges are a significant barrier when using machine learning tools for river basin pollution forecasting. Limited data availability, inconsistent sampling methods, and data handling with large high-dimensional volumes have been the most significant concerns in addressing such issues. They would facilitate improving predictive models, with further reliability and accuracy outcomes that could enable proper environmental management.

One of the main obstacles to predicting pollution in river basins is the non-existence of historical data. Many regions have not yet established long-term environmental monitoring systems, leading to a critical lack in the availability of comprehensive pollution data. Without proper historical data, it becomes difficult to train machine learning models. These models rely largely on extensive datasets to find patterns for accurate predictions. If data does not exist, then it is likely that results will be unreliable. It is said that such a hurdle can only be surmounted through the development of frameworks for data-sharing collaboration between the government,

research institutions, and other environmental organizations. Under such frameworks, diverse data from various outlets can be pooled and made easily accessible to everybody involved for better quantity as well as the quality of available information in modeling (Adamasi, et al., 2020).

Inconsistent sampling techniques can also present significant challenges when predicting pollution trends. Methods of sampling, rituals, identity of sources of sampling, as well as frequency of sampling all influence pollution data collection, thereby presenting biases and gaps in data coverage. Some examples of these include the case where irregularity of sampling of pollution sites is unable to directly reflect environmental conditions of the river basin or when sampling sites are selected in a non-representative way. It could also get challenging to develop models for robust predictions along with pollution trends over time. Standardization of protocols for sampling and uniformity in practices of data collection will go a long way in yielding reliable data along with prediction accuracy (Kumar and Singh, 2020).

Pollution datasets for river basins have been composed of several variables such as water quality parameters, meteorological data, land use, and pollution sources. This high dimensionality complicates the modeling process since every one of the machine-learning algorithms has to use several interrelated parameters for processing data peculiar to its analysis. Noise models can be permeated with irrelevant or redundant variables thus leading to overfitting or inefficient computational modeling. For example, such aspects are tackled through the use of feature selection and dimensionality reduction techniques that include having the principal component analysis (PCA). However, even with these features, high dimensional datasets are still a severe problem of computation when one's data is large-scale river-basin models in Li et al. (2019).

2.7.2 Model Challenges

The application of machine learning (ML) models in river basin pollution prediction is riddled with challenges concerning performance, interpretability, and integration. It is, therefore, advisable to address all of these challenges to make the

models both accurate and useful for decision-making. Such challenges of a model include overfit/underfit, an explanation of results, and hybridization between hydrological models and ML models.

One of the greatest challenges in machine learning is balancing between overfitting and underfitting. Overfitting occurs when a model becomes too complex and begins to learn noise or spurious details in the training data, making it generalize poorly on new, unseen data. Its performance measures, during the training of a model, seem quite inflated, but the model performs poorly in real life. In contrast, underfitting happens when the model is too simple to capture the inherent relationships present in the dataset; thus, the model makes inaccurate predictions. That is the challenge that machine learning professionals face-the task of building dependable models such that they end up perfectly between overfitting and underfitting. Regularization techniques, cross-validation, and hyperparameter tuning are some of the approaches employed to address the issue (Zhou et al., 2021).

Whereas in machine learning models, it is another challenge that is obviously related to explainability. It is more pronounced in models many, like deep learning. They are phenomenally called "black boxes" since they can deliver predictions while manipulating the input information without exposing how it does so. This nontransparent characteristic can make one lose trust in such models, especially when it comes to decision-makers who require a clear picture of how conclusions have been drawn. Interpretability in models is especially important for environmental applications where various stakeholders need to know how predictions are associated with initiating actions and developing policies. Research on explainable AI or XAI addresses this also by developing methods that seek to shed light on the poorly understood decision-making processes of complex models (Ahmed et al., 2020).

Integrating machine learning models into traditional hydrological models can greatly improve the accuracy and robustness of pollution prediction. With the help of hydrological models, rain and runoff, and stream flow considerations-all of which are significant for knowing the transport and fate of pollutants in river basins-are accounted for. By including these important dynamic environmental input variables in the machine learning model, it is possible to develop much more sophisticated

predictions which consider both pollution and natural processes affecting water quality. Unfortunately, this physically-based model does not allow much flexibility concerning data compatibility, model complexity, or computational efficiency: a sore point in developing hybrid models in which the benefits of machine learning are combined with those of traditional hydrological modeling mostly will find the solutions to providing much more accurate predictions of pollution dynamics within river basins (Zhou et al., 2021).

2.7.3 Future Directions

Future directions in the field of ML for predicting pollution incidentally make it promising to better model accuracy, interpretability, and practicality. Improved integration of multi-source data, development of explainable models, and new strategies such as federated learning would greatly stabilize the foundation to improve the reliability of environmental prediction models.

Emerging technology such as, Internet of Things (IoT) sensors, and real-time data analytics increases the potential of improving the accuracy of pollution predictions so that it can enable more responsive environmental management. Liu et al. (2022b) point out that the approaches currently being envisaged use these developments within existing modeling methods for covering pollution problems in a completer and more real-time, data-driven manner. Future research should address the aforementioned topics to eliminate current problems and hence, enable improved decisions in river basin management and pollution control.

The introduction of advanced artificial intelligence models such as transformers and generative adversarial networks into pollution prediction models over river basins would have the potential to revolutionize the whole process of predicting pollution. Transformers originated in the natural processing language domain and have been valuable to time series forecasting in the ability to manage long-distance dependence for long exposure séances. This is important for environmental predictions regarding pollutant levels that follow quite complicated temporal patterns as a result of seasonal changes, weather effects, and other human influences. By using transformers, they

would thus be able to improve pollution prediction accuracy and robustness as they can capture very minor sub-observances found in larger sequential data sets found in historical models (Vaswani et al., 2017).

Generative adversarial networks (GANs) are ideal tools for pollution prediction. They involve two neural networks, namely the generator and the discriminator, which create synthetic data imitating real-world patterns through cooperation. Such networks can be trained to produce synthetic pollution data from locations or time periods where little or no data exist in cases of not having any historical data. They can be used for improving the data gaps and enhancing model training. They generate more diverse and representative data for generalized models, which can perform better on different pollution scenarios (Goodfellow et al., 2014). All such advanced artificial intelligence models can contribute significantly to the enhancement of the predictive capacity of pollution models and hence improve the possible environment-friendly management measures.

Integrating real-time monitoring data from IoT devices in pollution forecasting models can greatly increase their accuracy and timeliness. For instance, sensor nodes embedded into the river ecosystem and floating on a water body are designed to generate continuous high-resolution data on several environmental parameters such as water quality, pollutant concentrations, temperature and flow rates. Real-time data becomes essential to capture the dynamism of pollution events, which are to a great extent affected by various factors such as weather patterns, human activities, and seasons. Feeds from these sources into predictive models can help researchers improve model training with more accurate and up-to-date predictions of pollution levels in river basins (Xu et al., 2020).

Real-time data provides models with constant updating, allowing them to adapt to new trends and evolving parameters. Such adaptabilities make the models even more reactive to instantaneous pollution events like those from industrial outfalls or severe weather events and help the decision-maker take appropriate actions in time. The merger of IoT with the machine learning models can even provide a basis for developing early warning systems that would alert the authorities of expected pollution threats allowing them to set up mitigation measures on time before the pollution

intensity increases. The advancements in IoT technology will only continue producing huge volumes of fine-grained data, yielding more capacity for pollution models to predict events, thus boosting river basin management efforts (Gao et al. 2021).

Integrating remote sensing data and satellite images into pollution prediction models can provide a huge enhancement in accuracy by providing spatial information about environmental parameters. Remote sensing technology has the potential to provide large area, high-resolution data on land uses, vegetation covers, water bodies, and pollution hot spots in very vast extent areas. This could be said during river basin management when a holistic picture of pollution could be established for very wide stretches and sometimes more remote areas. Analyzing satellite pictures to detect land use changes, such as urbanization or deforestation, and the amount of green vegetation is valuable because these changes and the health of vegetation are very significant factors affecting pollution levels in rivers. Moreover, remote sensing data can give insight into water quality by detecting some factors of turbidity, surface temperature, and chlorophyll that are mostly associated with pollution occurrences (Schwalm et al., 2020).

Considered as inputs of the machine learning models, such information would render holistic predictions that are also spatio-temporally accurate. For example, satellite-derived water quality indices would remedy the gap in field observations in areas where monitoring stations are few or non-existent. It can enable models that predict temporal-spatial patterns of pollution as all potential pollution hot-spot areas, land-use changes, or efficiencies for mitigation resource allocation. Besides, remote sensing data could validate model outputs, allowing improvement and greater accuracy in models (Tao et al., 2021). Newer remote sensing technologies could further extend the reach of prediction and monitoring of pollution events and thus transform this pollution into one of the most interesting applications of environmental indicators.

Globalizing pollution prediction models create a platform for identifying and understanding pollution patterns and trends in different geographical areas. This involves translating the models and methodologies, which are developed for site-specific river basins, to models that are relevant and applicable to global river systems. With such a system, researchers will be in a better position to ascertain from continent

to continent and from ecosystem to ecosystem how pollution dynamics differ. More importantly, such global insights can lead to identifying many common pollution problems, understanding their transboundary effects, and gauging their effects for designing environmental policies and regulations internationally (Vörösmarty et al., 2010). Scaling models account for the specific characteristics of the various river systems, such as the climate, land use, socio-economic conditions, and levels of industrialization, all of which play an important role in determining the pollution levels and their distributions.

Integrating and collating large-scale datasets such as remote sensing data, IoT sensor networks, and historical pollution datasets is significant to make possible the real establishment of models that can deal with the complexity and diversity of global rivers. It can be configured between machine learning techniques and this large body of data allowing better predictions at global levels. They can also be used to assess the changes impacts of climate change, population growth, and urbanization on river pollution with respect to water quality and future ecosystem health (Seitzinger et al., 2010). Yet, the models will face very serious challenges such as data availability, model calibration, and high computational power. Although these stages have major limitations in today's world, the aspect of being able to formulate solutions related to global pollution issues and informing scale-based large-essay integrated environmental management approaches stays at the forefront of what is necessary for future research.

2.8 Research Gap

The predictive modeling of pollution in river basins holds numerous key challenges that restrict the effectiveness and applicability of current methodologies. One such gap is that of data quality because most studies are based on data with missing, incomplete, or unreliable measurements, which undermine the performance of machine learning models. Further, sophisticated machine learning algorithms like neural networks and deep learning mostly lack interpretability, which makes understanding their predictions very difficult for stakeholders, researchers, and policymakers, hitherto limiting the scope of effective usage of their predictions in the

decision-making processes. Another gap is the poor integration of machine learning models with traditional hydrological models; thus, such venturing into the development of holistic approaches that consider both data-driven insights and hydrological dynamics remains limited. Furthermore, most of the existing research works are concentrated on isolated case studies, denying the generalization of models to diverse river systems and geographical regions. Such limitations provide a good reason for more robust, accurate, and flexible predictive modeling frameworks.

Thus, this research intends to take predictive models for river basin pollution to a new level in terms of accuracy, interpretability, and applicability. It will involve the development of advanced machine learning techniques with the consideration that the models will be easily understood by stakeholders. This study, furthermore, wants to combine machine learning approaches with traditional hydrological models to build an all-encompassing system that identifies data-driven patterns and physical processes driving pollution dynamics. Moreover, improving preprocessing techniques and better data sources will enhance the accuracy of models. This research would provide generalized frameworks that can be adapted into any river system and allow practical usage in different parts of the world. These issues are addressed, and the study would greatly contribute to environmental management since it places tools into the hands of policy-makers and stakeholders in making informed decisions on protecting and sustaining important freshwater resources.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

It provides an exhaustive methodological framework to develop and validate predictive models to estimate pollution in river basins by using machine learning techniques. In other words, all data on the environment, water quality, and socioeconomic aspects would be harnessed to develop accurate and correctly actionable models for guiding mitigation strategies. In this way, one would not only systematically handle data but also develop specifically robust model selection so that one could use them in more realistic applications. The research process, such as how such a study begins and how the results and models are evaluated, is discussed more. In this chapter, the dataset used and its performance measurement will be introduced and illustrated.

3.2 Research Framework

The research framework for this study is structured to provide a comprehensive methodological approach to addressing the complexities of pollution prediction. There were five phases of research to develop predictive models for pollution in river basins. Phase one is planning and initial study which contributed to problem formulation and background research. A milestone of an overview of points of interest can be identified and enable an insight into the whole project. Besides that, data preparation fell into phase two in which a cleaned dataset that was ready for further analysis was well-prepared. Furthermore, phase three is to retrieve the relevant features from the pre-processed dataset. In this phase, the underlying pattern of the dataset can be identified by machine learning model by LSTM, from pre-processed data. Additionally, a random forest model will be implemented into the retrieved features to classify the data based on their similarities. In this case, a milestone for the source of river basin

pollution can be illustrated. Lastly, R-squared was applied for model evaluation. The relationship between sources and river basin pollution can be visualized in this phase. Figure 3.1 illustrates the overall research framework. This will elaborate on every phase.

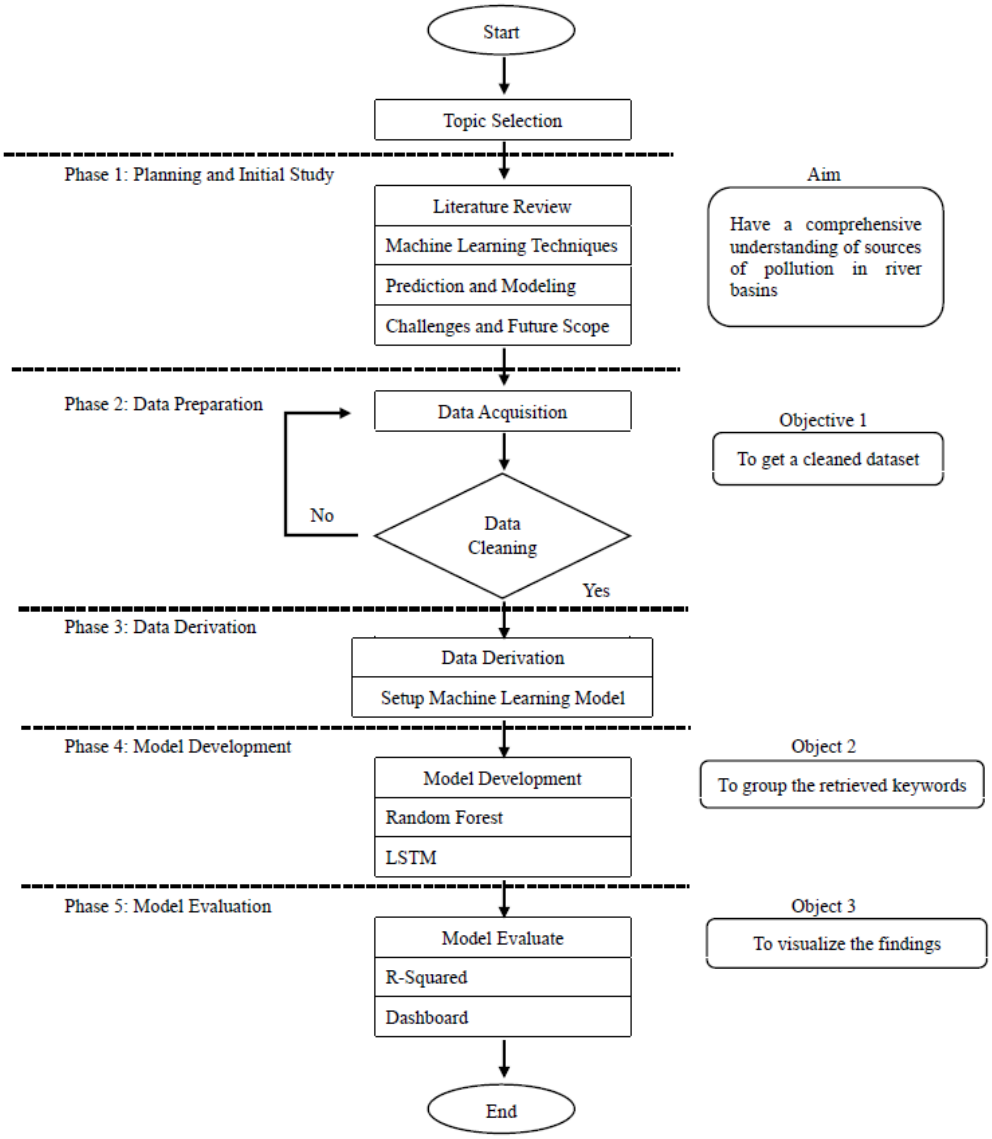


Figure 3.1: Overall research framework

3.2.1 Planning and Initial Study

This phase is laying the groundwork for research by several significant strands. The major component is a detailed literature review to assess studies related to the pollution of river basins, predictive modeling methods, and relevant data sources.

Machine learning algorithms will also be looked into for potential suitability in time-series and spatial data to identify existing research lacunae and to guide the choice of appropriate models as well as courses of action in data collection. Then, that phase is problem definition and scope, which means the precise problem statement with regard to the research. This includes the identification of a target river basin or basins and pollutants of interest, for example, specific chemical compounds or biological indicators, and also definitions in the temporal and spatial scope of the study.

Research question development aligns with the SMART criteria which stand for specific, measurable, achievable, relevant, and time-bound. The questions will address the investigation's general aspects, such as the factors that have the most influence on pollution levels within the target river basins or the performance prediction of the different machine learning algorithms considered. Moreover, whether the developed models will generalize to unseen data and the practical implication of these predictive models for environmental management and policy-making are significant questions regarding this investigation. This approach gives a solid base for meaningful and powerful research.

3.2.2 Data Collection and Preparation

Identifying and using several sources of relevant data will be required for extensive coverage of the parameters concerning pollution and the factors that affect them. The historical water quality monitoring data will be collected from monitoring stations with the target river basins, covering the parameters: biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), pH, turbidity, nutrients, and concentrations of heavy metals.

These include environmental data that should include factors such as rainfall patterns, temperature fluctuations, river flow rates, land use characteristics, and agricultural practices that may affect water quality. Socioeconomic data are equally important to include population density, industrial activities, capacity of wastewater treatment plants, and extent of agricultural activities within the basin. This integration

of diverse approaches will result in the creation of a substantial and multi-faceted dataset that serves the analysis and predictive model for water basin pollution.

Data Preprocessing is the very first stage in preparing the data collected to be fit into machine learning algorithms. This step will include taking care of the missing values. Where appropriate techniques like mean imputation or regression imputation would have to be used to fulfill the dataset for it to be usable. Then, data normalization in itself is part of the process whereby the data set will be scaled from its initial range to a common range that helps in improving performance and convergence in machine learning.

Feature engineering is followed, in which new features are created either using existing variables or through the addition of domain knowledge to improve the prediction accuracy of models. Finally, the data sets are split into training, validation, and test sets. This is done so that systematic model building and evaluation can take place, which enhances performance assessment and reduces the chances of overfitting.

3.2.3 Data Derivation

This research data derivation is the extraction of useful information from already preprocessed data to base predictive modeling of pollution levels in the river basin. The process uses high-end language models like ChatGPT to derive insights and patterns to be incorporated into the machine-learning pipeline.

The connection between the research system and ChatGPT has indeed been established using the OpenAI-provided API. Here, the acquiring of an API key and utilizing this key to interface with the data processing framework to provide requests are the 2 most important activities.

Once the API starts producing results, the next logical step is to establish prompt instructions that set the stage for the language model to operate in the specific task-oriented direction defined by the research outcomes. These instructions would be well-crafted to analyze pre-processed river basin data and extract informative insights from

it. For example, prompts could instruct the model to identify pollutants present in water, determine seasonal patterns in that pollution, or even find areas with the most pollution, giving results in specified JSON formats.

The next step involves the processing of API responses that are formed in JSON according to the specification required by the output consumer for further analysis. This could be analysis in terms of grouping pollutants according to their source-for example, industrial, agricultural, or urban runoff-and how they affect river ecology. Such categorized findings allow the user to know more about pollution sources and their weight as severity.

Finally, the machine learning models use data that are derived through all of this predictive analysis. API structured outputs are integrated with pre-processed, such as meteorological and hydrological variables, into full datasets that inform up-to-the-moment inputs for training and validating predictive models on pollution predictions and environmental management.

The data derivation phase is the vital step that links raw data and works toward actionable insights in making sure that the machine learning models are fed relevant and context-rich data. Largely as a result of using LLMs, complex relationships and latent patterns in river basin data can also be revealed in a way that would otherwise be quite difficult to derive from traditional preprocessing methods alone. This would, thus, improve the predictivity of models in addition to creating a more accurate and efficient pollution prediction system.

3.2.4 Model Selection and Development

The process for selecting and developing a model consists of identifying and applying appropriate machine learning algorithms to predict pollution levels on the data characteristics and the study's objectives. Different algorithms may be selected for a task, such as regression models such as linear regression, support vector regression (SVR), random forest regression, gradient boosting machines (GBM), and extreme gradient boosting (XGBoost). Neural networks, such as multilayer

perceptrons (MLP), recurrent neural networks (RNN), and convolutional neural networks (CNN), can also be employed, especially for capturing complex patterns in temporal or spatial data. Additionally, ensemble methods like bagging, boosting, and stacking leverage multiple models' strengths for improved predictive performance.

Once the algorithms are selected, they are trained using the prepared training data, optimizing parameters such as learning rates and regularization factors to enhance their performance. Model validation follows, where the trained models are assessed using a separate validation dataset to evaluate their accuracy and ability to generalize to unseen data. Key metrics for evaluation include R-squared (R^2), which measures the proportion of variance explained by the model, root mean squared error (RMSE), which quantifies the average magnitude of prediction errors, and mean absolute error (MAE), which calculates the average absolute difference between predicted and actual values. These metrics provide a comprehensive understanding of model performance and guide further refinements.

3.2.5 Model Evaluation and Interpretation

Model evaluation and interpretation involve assessing the performance of various machine learning models to identify the best one for predicting pollution levels. This is achieved by comparing the models using chosen evaluation metrics, such as R-squared, RMSE, and MAE, to determine which model delivers the highest accuracy and generalization ability.

Feature importance analysis is then conducted to understand the relative significance of different input features in influencing the predictions. Techniques like permutation feature importance, which involves randomly shuffling the values of a feature and observing changes in model performance, provide insights into how critical each feature is. Additionally, SHAP (SHapley Additive exPlanations) can be employed to explain the model's output by attributing contributions to individual features, offering a deeper understanding of their roles.

Model interpretability is also a key focus, as it helps uncover the relationships between input features and pollution levels. Visualization techniques, such as plotting predicted pollution levels against actual values, are used to assess model accuracy and identify patterns. Furthermore, analyzing feature interactions reveals how different variables work together to impact pollution levels, enhancing the interpretability and reliability of the model for practical applications.

3.2.6 Model Deployment and Application

Model deployment and application involve implementing the selected machine learning model for real-time prediction of pollution levels in the target river basins. This process ensures that the model is integrated into a practical system capable of generating timely and accurate predictions. A user-friendly interface is then developed to allow stakeholders to easily access model predictions and visualize results. This interface may include interactive dashboards, charts, or maps to make the outputs intuitive and actionable.

The findings from the research are communicated effectively to stakeholders, such as policymakers, environmental agencies, and local communities. This ensures that the insights derived from the model inform decision-making processes and support the development of targeted pollution mitigation strategies. By bridging the gap between complex data analysis and real-world application, this phase ensures the research has a tangible impact on environmental management efforts.

3.3 Dataset

The river basin pollution review dataset was obtained from the Department of Statistics Malaysia (DOSM) repository. It consists of 10,840,500 rows of data representing the 150 river basins for three primary water pollution indicators monitored by the DOE presented in this dataset which is Biochemical Oxygen Demand (BOD), which measures the amount of oxygen that microorganisms require to decompose organic matter in the water. High BOD levels indicate the presence of a

large amount of organic pollution, Oxygen-depleting elements in the water such as ammoniacal nitrogen (NH -N), which measures ammonia present in water primarily resulting from agricultural runoff, sewage, and industrial discharges, would harm aquatic life. The ammoniacal nitrogen level plays a toxic effect on fish and aquatic organisms and can cause an oxygen depletion due to the ammoniacal nitrogen being broken down into its components and Suspended Solids (SS) which consist of all very small particles in the water including soil, silt, organic debris, and industrial waste. Suspended solids in high concentrations can decrease the clarity of water, inhibit sunlight reaching aquatic plants, and block the gills of fishes. According to Water Quality Index (WQI) values for the three indicators, river basins are classified as clean, slightly polluted or polluted.

Table 3.1: Riven Basins Pollution Dataset

Date	**Basins Monitored**	**Pollution Indicator**	**Pollution Status**	**Number of Basins**	**Proportion**
2008-01-01	143	nh3n	polluted	33	23.1
2009-01-01	143	nh3n	polluted	40	28
2010-01-01	143	nh3n	polluted	42	29.4
2011-01-01	140	nh3n	polluted	35	25
2012-01-01	140	nh3n	polluted	38	27.1

3.4 Performance Measurement

R-squared (R^2) is a metric that tells us how well our machine-learning model predicts the levels of pollution in river basins. This reveals the extent to which the changes in pollution levels can be explained by the components we included in the

model, such as weather, industrial waste, and runoff from agriculture. For example, if the R^2 value is 0.85, so in this model, 85% variability in pollution levels occurs because of factors. The remaining 15 % is possible because of other factors which are not included in the model or random noise.

The first stage of implementation of R^2 involves collecting data on pollution levels and other factors related to them. Thereafter, the machine learning model is provided with this data for training purposes. As a result, the machine should now be able to make predictions regarding pollution values. The next step is to compare the predictions made by the model with the actual values of pollution level readings recorded so one may calculate the accuracy. R^2 will be higher the closer the predicted values are to those measured. Thus, one's R^2 will be very near to 1 if the predictions from the model are near to actual reality, indicating a good prediction. However, if R^2 is close to 1-the predictions will be far from reality- that could indicate poor prediction performance by the fitted model.

R^2 would help in this research to determine how reliable our model would be in predicting levels of pollution based on the data that we provide. A high R^2 would mean that the model is well able to capture important relationships between the data and pollution, which allows for the model to be credible in decision-making or future predictions. A low R^2 would suggest the model may need further improvement or that more factors should be considered.

3.5 Future Research Directions

Future research in this area can focus on several of the most advanced approaches toward predictive modeling of pollution in river basins. A promising direction could be deep learning approaches, for example, convolutional neural networks or the use of long short-term memory (LSTM) networks to capture complex temporal and spatial characteristics that may escape traditional model structures. These approaches seem to be most useful for revealing pollution level changes in time and space throughout the river basin. Other hybrid models, combining physical approaches, such as hydrological simulations, with statistical and machine learning

techniques, may further broaden our understanding of pollution dynamics while improving prediction accuracy.

There is another area of research that will include the integration of multiple data sources to improve the performance of the model. 'IoT' sensors installed in and around river basins can stream real-time data combined with satellite imagery and provide a real-time flow of information regarding environmental conditions. The challenges of data sparsity can be overcome by transfer learning or synthetic data generation. Transfer learning facilitates model generalization to apparently related fields while the synthetic data generation procedure can help create more data slices to supplement training activities with limited real data.

Transparency was the last ranking critical for building trust among stakeholders for model development and expected results. Future research has to improve model interpretability using visualization tools and explainability methods that would enhance understanding of prediction-making; and construct trust in the models, hence regarded as credible, for their predictions. The contract requires additional procedures to broaden the confidence of stakeholders, namely by validating them against independent data sets providing external proof for the actual efficacy in the real world of the model.

3.6 Conclusion

It was indeed a robust framework based on prediction modeling of river basin pollution using machine learning. Addressing every step in research guarantees the scientific soundness, practical relevance, and usefulness of action. Future studies would largely be based on advanced technologies and methodologies for improved model accuracy and applicability.

CHAPTER 4

INITIAL RESULTS

4.1 Introduction

This chapter discusses the initial results of applying various machine learning models to predict water quality in river basins. It probably involves collecting data on physical and chemical parameters of water quality, such as Total Suspended Solids (TSS), Ammoniacal Nitrogen (NH3N), and Biochemical Oxygen Demand (BOD). Decision Trees, Artificial Neural Networks, K-nearest neighbors, Naïve Bayes, Support Vector Machine, Random Forest, and Gradient Boosting are then used as the feature engine for machine learning algorithms. The initial results suggested that Gradient Boosting achieved the highest accuracy, sensitivity, and f-measure when predicting water quality and hence, this model seemed to have the best predictive power. With the most significant features highlighted for the model, it will be able to provide the significant parameters that most influence predictions with respect to water quality.

4.2 Data Visualizations

In general, the river quality in 2023 has slightly deteriorated as compared to the previous year, based on the Water Quality Index. In this country, every year, there are 1,353 manual river water quality monitoring stations with a coverage area of 672 rivers. This Department of Environment, Malaysia will continue its monitoring to support the strategies, programs, and activities concerned for the sustainable management of the environment effectively. River water quality monitoring is done to determine the status of river water quality and to detect changes in river water quality. Water samples were collected from designated stations for in-situ measurement and sent to the laboratory as well, for analysis aimed at determining the criteria based on the sciences of physic-chemical and biological. The WQI is used to indicate the level

of pollution and the corresponding suitability in terms of water use according to the National Water Quality Standards for Malaysia (NWQS). NWQS is an ambient standard to protect aquatic biodiversity and it is also used as a benchmark for the setting of uses for certain rivers.

WQI is an index range from 0 to 100 where the range is divided into three (3) categories clean, slightly polluted, and polluted. The index is derived based on six (6) parameters, which are dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (AN), suspended solids (SS), and pH. Reporting water quality status is reported in two (2) approaches: River water quality status by the river and River water quality status by station. Initial result from the Environmental Quality Report (EQR) 2023, out of the 672 rivers monitored, 25 (4 percent) were polluted, 161 (24 percent) were slightly polluted, and 486 (72 percent) showed good water quality. In 2022, there were six polluted rivers in Malaysia, a decrease from seven polluted rivers in the previous year. The lowest number of polluted rivers in Malaysia was in 2015, with 5 rivers in total. In the same year, there were a greater number of clean rivers in Malaysia.

For water quality status reporting by river, shown in Figure 4.1 the WQI determination is based on the median calculation where the median WQI for the entire data observed at each station to obtain the river WQI. The reporting of water quality status by river is a reflection of the quality status of the entire river based on the number of stations monitored.

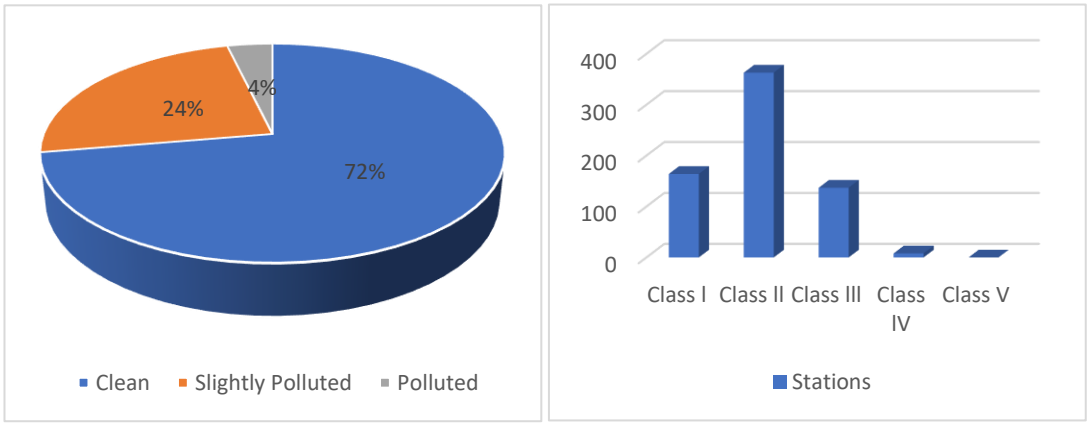


Figure 4.1: Water quality status by river

For water quality status reporting by station, the WQI determination is based on the median calculation of six (6) WQI at the station. The reporting of water quality status by the station is a reflection of the quality status at the station only and is shown in Figure 4.2.

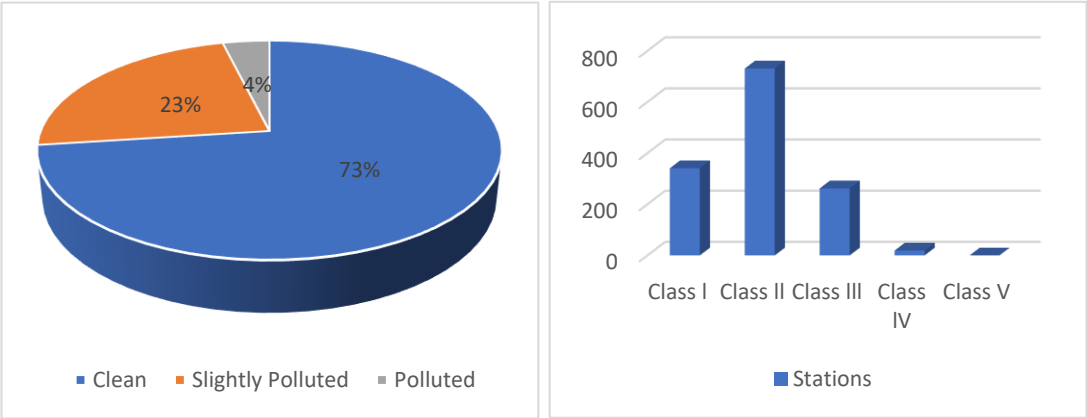


Figure 4.2: Water quality status by station

The water quality status of a total of 672 rivers was monitored every year. Out of the 672 rivers monitored, the water quality of 486 (72%) rivers was indicated as clean, 161 (24%) rivers were indicated as slightly polluted, and 25 (4%) rivers were indicated as polluted. The trend of the monitored river water quality is shown in Figure 4.3. For WQI classification, as many as 164 (25%) rivers are in Class I, 363 (54%) rivers are in Class II, 137 (20%) rivers are in Class III, and eight (8) (1%) rivers are in Class IV.

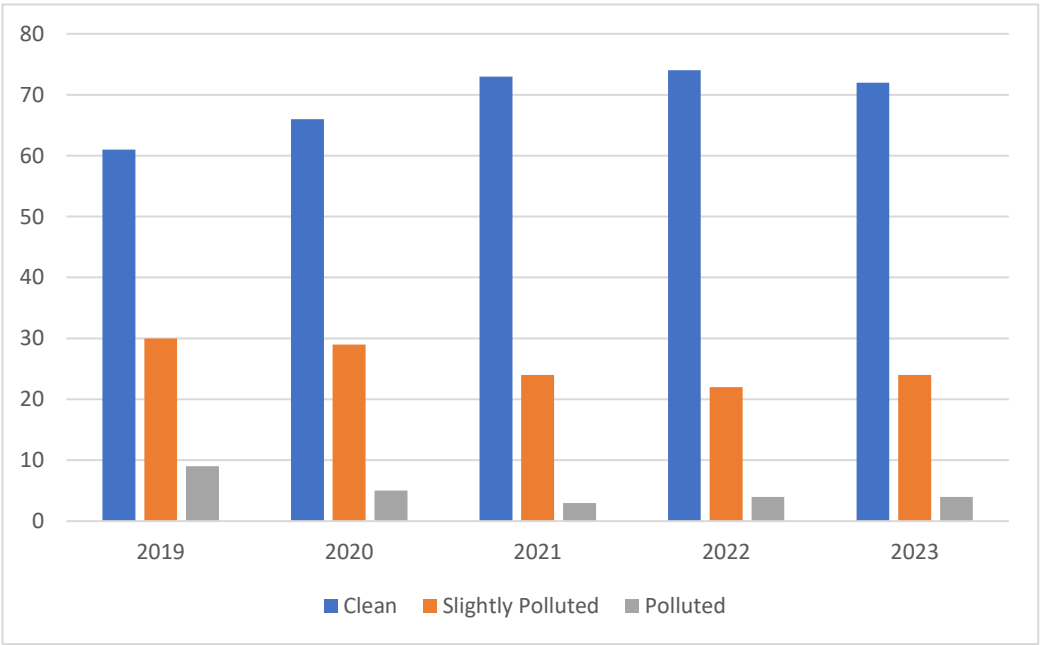


Figure 4.3: River Water Quality Trend from 2019 until 2023

The main indicators were considered to determine river water quality based on the parameters of BOD, AN, and SS. The cause of deterioration of water quality is due to the discharge of the pollution load either from a point source or from a nonpoint source. Point sources can be described as pollution sources with specific identifiable discharge points that are unchanged over time. Sectors such as industry, livestock, and sewage treatment systems fall under this category. Meanwhile, the nonpoint sources such as agricultural activities, earthworks, mining, and sullage (domestic wastewater other than sewage such as kitchen and bathroom wastewater) do not have specific identifiable discharge points and the locations are varied over time. This makes it difficult to estimate the amount of released pollution loads.

4.2.1 Stores Data

The details of dataset construction and preparation steps for the first research objective. Table 4.1: Longitudinal study-the data was collected from 144 river basins over some time, starting from January 1, 2019, and ending on December 31, 2024. The main focus will be on monitoring pollution levels based on different indicators, such as BOD5, NH3N, and SS. The dataset also contained the date of monitoring, status of pollution, and proportion. The above analysis indicated a remarkable improvement in the water quality of river basins, particularly for BOD5 and SS. This study will cover predictive modeling of the Sungai Johor River basin pollution.

Table 4.1: River Basin Pollution Monitoring Dataset

	Date	Basins Monitored	Pollution Indicator	Pollution Status	Number of basins	Proportion
0	1/1/2019	144	bod5	clean	53	36.80555556
1	1/1/2019	144	bod5	slightly__polluted	39	27.08333333
2	1/1/2019	144	bod5	polluted	52	36.11111111
3	1/1/2019	144	nh3n	clean	12	8.333333333
4	1/1/2019	144	nh3n	slightly__polluted	74	51.38888889
5	1/1/2019	144	nh3n	polluted	58	40.27777778
6	1/1/2019	144	ss	clean	78	54.16666667
7	1/1/2019	144	ss	slightly__polluted	36	25
8	1/1/2019	144	ss	polluted	30	20.83333333
..	…	…	…	…	…	…
16417	31/12/2024	144	bod5	clean	124	86.11111111
16418	31/12/2024	144	bod5	slightly__polluted	11	7.638888889
16419	31/12/2024	144	bod5	polluted	9	6.25
16420	31/12/2024	144	nh3n	clean	58	40.27777778
16421	31/12/2024	144	nh3n	slightly__polluted	37	25.69444444
16422	31/12/2024	144	nh3n	polluted	49	34.02777778
16423	31/12/2024	144	ss	clean	127	88.19444444
16424	31/12/2024	144	ss	slightly__polluted	7	4.861111111
16425	31/12/2024	144	ss	polluted	10	6.944444444

[16425 rows x 6 columns]

Figure 4.4 Number of main river basins in states of Malaysia. The x-axis is the chart represents states while y-axis represents total number of main river basins. Each of the bars in the chart has been labeled with the corresponding state and the number of its river basins. From the chart, Sabah has the highest number of main river basins, at 75, followed by Sarawak, which has 40. Johor has 20 main river basins, while Pahang and Terengganu each have 12. Perak is shown to have 11 main river basins, and Negeri Sembilan has 8. Kedah is listed with 7 main river basins, Melaka and Selangor each have 6, and Pulau Pinang has 5. Kelantan has 4 main river basins, WP Kuala Lumpur has 3, and Perlis has the least, with only 1 main river basin. The above information provides good insight into how the main river basins are distributed across various states in Malaysia. Such information is useful in studies related to geography, environment, and resource management.

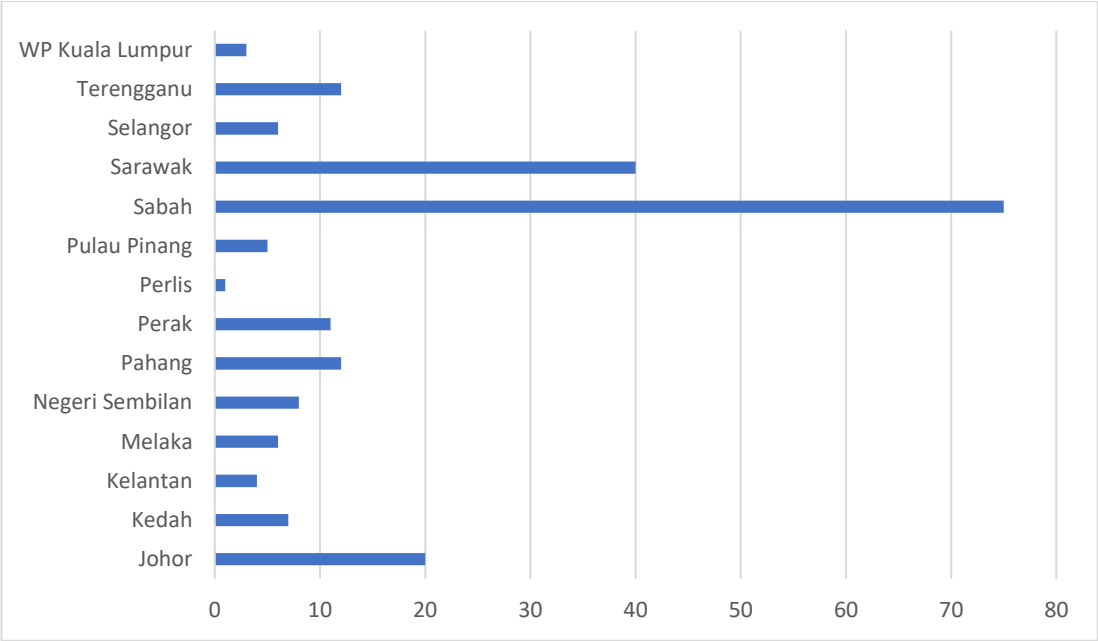


Figure 4.4: Main River Basin by State

Figure 4.5 Summary of water quality assessment for thirteen states of Malaysia. Each state is represented by a set of three bars representing the number of river basins falling into the three categories of clean, slightly polluted, and polluted. The height of each bar directly gives the number of river basins in that category of pollution for the corresponding state. Instantly, the chart visually underlines those states that have clean river basins-for example, Johor, Pahang, and Sabah-and the ones with higher shares of polluted or slightly polluted basins, like Pulau Pinang and WP Kuala Lumpur. Such coloring for each pollution level enhances clarity in visual form, and thus it presents faster comparisons between states.

However, the simplicity of Figure 4.5 limits the depth of the analysis that it can provide. For example, the particular water quality indicators that were used to classify the pollution status of the river basins are not specified (e.g., BOD, dissolved oxygen, nutrient levels). Also, the period over which data has been collected is not specified, so one cannot know whether the data presented reflects conditions of a snapshot or longer-term ones. The presentation of information without any indication of methodologies adopted for collecting data-sampling frequency, locations, and analytical techniques-suggests issues with the reliability and interpretability of the results. In this way, although the chart provides an overview, further analysis would in any case require more contextual detail to reach comprehensive conclusions.

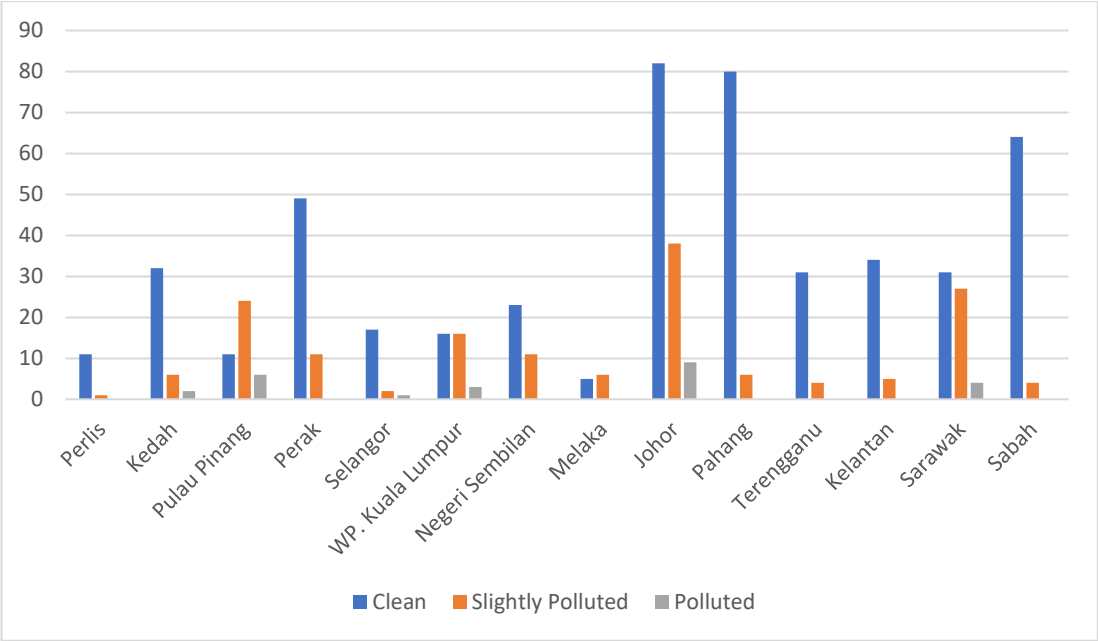


Figure 4.5: Pollution in River Basins by State

4.2.2 Item Data

Table 4.2 gives an overview of the water quality for different states of Malaysia. The first column, "States," refers to the geographical areas within which the respective assessments were done. The column "Clean" represents those river basins in a state that had been classified as "clean," or where the water quality is good. The slightly polluted column represents the number of river basins corresponding to slightly polluted, which is understood to be a state of moderate levels of contamination. The last column, polluted is the number of river basins that correspond to the category of polluted, meaning serious contamination of water.

The data indicates, based on general observation, that Johor and Pahang contain the largest number of clean river basins, which means these states generally possess good water quality. Sabah also reflects good water quality conditions with a high number of clean basins and very few polluted basins. Pulau Pinang presents the highest number of slightly polluted basins and hence could be considered a potential risk regarding water quality and WP. It is expected that in Kuala Lumpur, high counts for both slightly polluted and polluted basins may suggest a probable impact due to urbanization and industrial activities.

Table 4.2: River Basin Pollution Monitoring by State

State	Clean	Slightly Polluted	Polluted
Perlis	11	1	0
Kedah	32	6	2
Pulau Pinang	11	24	6
Perak	49	11	0
Selangor	17	2	1
WP. Kuala Lumpur	16	16	3
Negeri Sembilan	23	11	0
Melaka	5	6	0
Johor	82	38	9
Pahang	80	6	0
Terengganu	31	4	0
Kelantan	34	5	0
Sarawak	31	27	4
Sabah	64	4	0

Figure 4.6 shows that Johor has the highest proportion of 36% when looking at the river basins that are polluted, signifying the high level of pollution the state has. Pulau Pinang contributes 24%, equally considered as a high contributing to pollution. Sarawak and WP Kuala Lumpur are contributing fairly, at 16% and 12%, respectively. Selangor and Kedah thus appear to have the least problem with only 4% and 8% of the basins polluted, respectively.

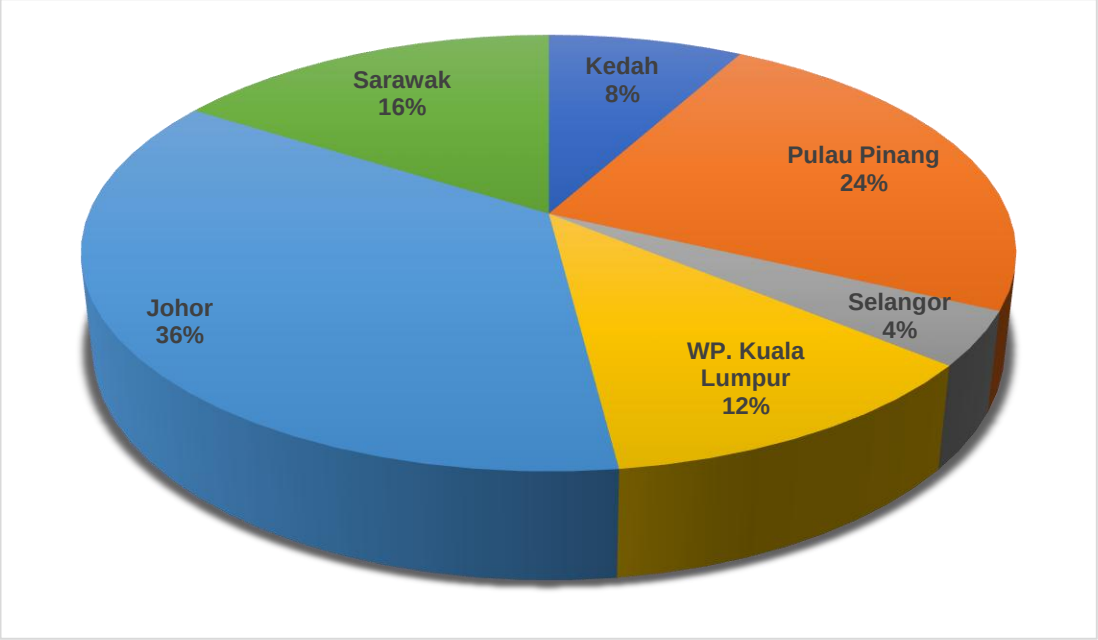


Figure 4.6: Polluted River Basin by State

Waste from manufacturing industries involved in the production of electronics, textiles, and food contaminates the rivers in Johor and Pulau Pinang because of improper waste disposal, which releases hazardous chemicals, heavy metals, and solid wastes into the water. Intensive agriculture, including large-scale animal husbandry and plantations, also pollute rivers through the addition of chemical fertilizers and pesticides and wastes from animals to the water system. This encourages eutrophication-a process reducing oxygen levels in the water and highly detrimental to aquatic life. For Pulau Pinang, tourism adds to pollution through inappropriate household and food waste disposal and possible oil or chemical spills from activities in the water.

Other contributors include construction sites through soil erosion and sediment runoff, as well as chemicals such as cement and paint. Finally, transportation includes shipping and vehicular traffic around rivers introduce oil spills, among other wastes into the water. Again, this is by no means an exhaustive list, and at most locations of river pollution, the actual pollution sources are caused by a combination of some or all of these elements and several others. Specific research will be required in finding out precisely the sources of pollution in each locality.

4.2.3 Merged Data

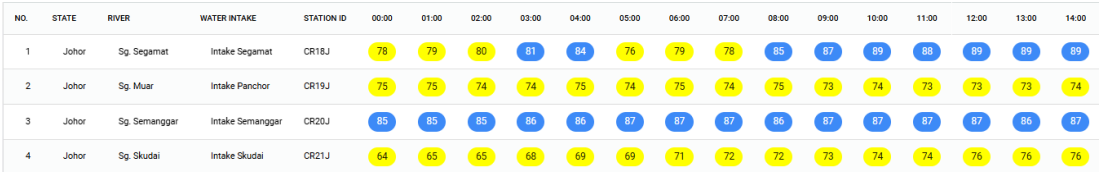
In this section, the manual river water quality index is concatenated by stacking these data frames horizontally as these data have the same format but with different values. Then, the Data frames are then merged with the manual river water quality index and continuous river water quality index on state, river, and station ID as shown in Figure 4.7 and Figure 4.8

NO.	STATE	RIVER	STATION ID	WQI/SAMPLING MONTH					
1	JOHOR	AIR BALOI	3JABL001	68 [September 2024]	74 [July 2024]	61 [May 2024]	74 [March 2024]	51 [January 2024]	58 [November 2023]
2	JOHOR	AIR BALOI	3JABL002	67 [September 2024]	60 [July 2024]	54 [May 2024]	54 [March 2024]	52 [January 2024]	53 [November 2023]
3	JOHOR	AIR BALOI	3JABL003	67 [September 2024]	73 [July 2024]	56 [May 2024]	61 [March 2024]	49 [January 2024]	54 [November 2023]
4	JOHOR	BENUT	3JBNT001	90 [September 2024]	82 [July 2024]	91 [May 2024]	96 [March 2024]	90 [January 2024]	89 [November 2023]
5	JOHOR	BENUT	3JBNT002	79 [September 2024]	88 [July 2024]	80 [May 2024]	88 [March 2024]	89 [January 2024]	87 [November 2023]

Figure 4.7: Manual River Water Quality Index

Figure 4.7 shows the Water Quality Index (WQI) sampling data for different stations in Johor, Malaysia. The table lists the station number, state, river, station ID, and WQI values for various months. Each month's WQI value is color-coded within the table. There are no questions to answer within the provided image; it's purely a data presentation.

Meanwhile, Figure 4.8 shows a table of hourly water quality readings from four different water intake points in Johor, Malaysia. The table includes the number, state, river, water intake location, station ID, and hourly water quality readings from 00:00 to 14:00. The readings appear to be some type of index, with values ranging from the low 60s to the high 80s. The values are color-coded, with yellow indicating lower values and blue indicating higher values. There are no questions to be answered in the image; it's simply a presentation of data.



NO.	STATE	RIVER	WATER INTAKE	STATION ID	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00
1	Johor	Sg. Segamat	Intake Segamat	CR18J	78	79	80	81	84	76	79	78	85	87	89	88	89	89	89
2	Johor	Sg. Muar	Intake Panchor	CR19J	75	75	74	74	75	74	75	74	75	73	74	73	73	73	74
3	Johor	Sg. Semanggar	Intake Semanggar	CR20J	85	85	85	86	86	87	87	87	86	87	87	87	87	86	87
4	Johor	Sg. Skudai	Intake Skudai	CR21J	64	65	65	68	69	69	71	72	72	73	74	74	76	76	76

Figure 4.8: Continuous River Water Quality Index

The information presented in Table 4.3, shows the average quality index for each state due to river basin pollution. It thus gives a proper overview of different states from the year 2019 until 2023 concerning river basin pollution. Within this table are states such as Perlis and Kedah, Johor having an average quality index for every year across this period. The color-coded values provide an easy-to-view visual indication of the level of pollution, ranging from green for good quality to red for poor quality. A "Status" column has been added to the table, summarizing each state as SP: Slightly Polluted, P: Polluted, or C: Clean based on data available for 2023. For example, in the year 2023, Perlis has the notation SP, which means slight pollution, whereas Johor has P, representing higher severity in terms of pollution.

The table highlights the variations in water quality over time, emphasizing areas that have either improved or deteriorated. For example, Perlis shows a generally

increasing trend in its quality index, moving from 71 in 2019 to 81 in 2023, indicating an improvement in water quality. In contrast, Johor has a low and rather consistent quality index, mostly varying within the 30s and 40s, indicating continuous pollution. This series of information is vital to monitor and ease the environmental challenges arising due to the pollution in river basins across various states.

Table 4.3: Average quality index for river basin pollution in every state

State	2019	2020	2021	2022	2023	Status
Perlis	71	75	76	74	81	SP
Kedah	64	70	77	76	73	SP
Pulau Pinang	66	65	63	56	41	P
Perak	85	88	91	86	86	C
Selangor	66	70	71	77	77	SP
WP. Kuala Lumpur	49	71	63	44	57	P
Negeri Sembilan	69	74	73	80	79	SP
Melaka	65	72	71	74	72	SP
Johor	36	29	33	39	33	P
Pahang	89	88	92	91	91	C
Terengganu	87	87	88	89	91	C
Kelantan	81	80	80	85	85	C
Sarawak	89	90	88	91	91	C
Sabah	73	74	80	69	70	SP

Figure 4.9 highlights the trend in Biochemical Oxygen Demand (BOD5) from 2000 to 2021. BOD5 is crucial because it measures the oxygen needed by microorganisms to decompose organic matter in water, with higher values indicating more pollution. From 2000 to 2014, there was a steady increase in BOD5 levels, reflecting a rise in water pollution. This period's increasing pollution could result from heightened industrial activities, urbanization, and insufficient wastewater treatment facilities. The upward trend underscores the escalating stress on water bodies due to human activities.

However, the trend in BOD5 levels decreased between 2014 and 2021, as represented by the graph, reflecting great improvements in water quality. It may be because of many reasons, such as a rise in the stringency of environmental laws, better wastewater treatment processes, and industrial practices with greener methods. This

downtrend indicates an effective implementation of pollution control measures and increased awareness of the need to protect this vital resource. This is a good omen for the recovery and gives testimony to effective targeted environmental intervention.

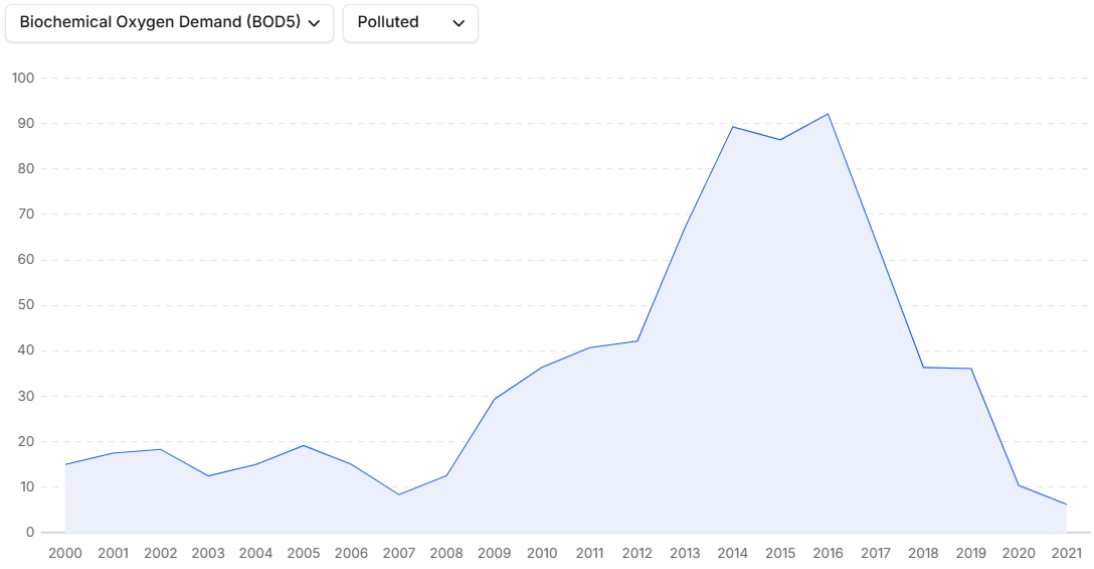


Figure 4.9: River Water Quality Stations Trend on *Biochemical Oxygen Demand (BOD)* Sub-Index

Figure 4.10 presents the trend of Ammoniacal Nitrogen (NH3-N) from the year 2000 to 2021. NH3-N is a certain type of nitrogen pollution usually discharged into water bodies through agricultural runoff, sewage, and industrial waste. Before 2019, NH3-N showed an increasing trend, which indicates a deteriorating situation in water pollution. The probable causes are increasing agricultural activities, urbanization, and inappropriate treatment of waste, leading to increased nitrogenous substances discharged into the water sources.

However, the graph for NH3-N levels showed a decline in the years 2020 and 2021, which is an improvement in water quality. This may be because of a number of factors that include more stringent pollution control regulations, improved wastewater treatment technologies, and changes in agricultural activities to more sustainable farming techniques. This effort seemed to be reducing nitrogen pollution in the water body and proved the efficacy of pinpointed environmental intervention and policies.

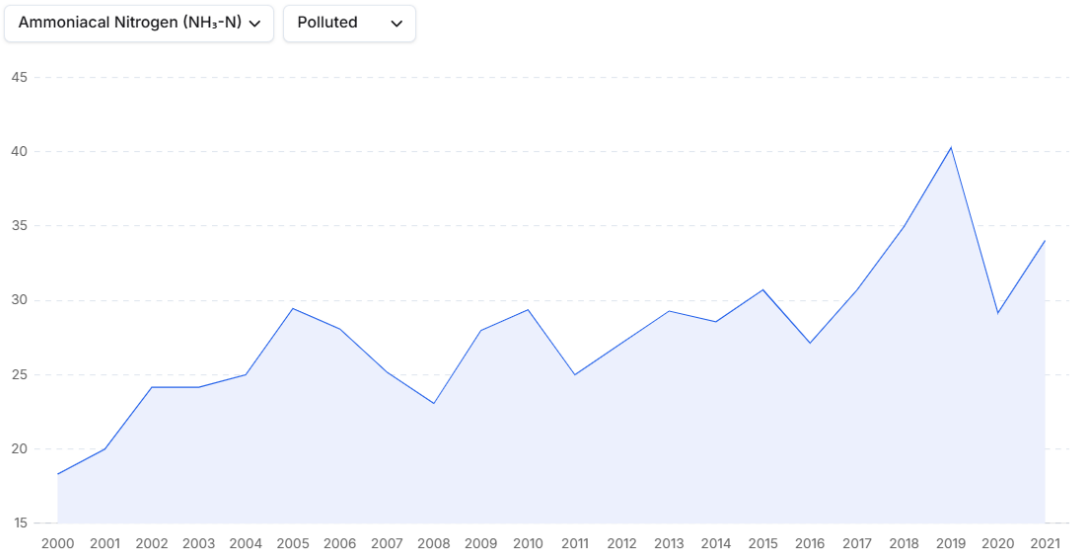


Figure 4.10: River Water Quality Stations Trend on Ammoniacal *Nitrogen (NH_3-N)* Sub-Index

Figure 4.11 presents the trends in Suspended Solids (SS) levels in water from 2000 to 2021. Suspended solids are small particles that float in water, generally comprising soil, organic matter, and man-made products from various industries. High levels of SS result in turbid water, not only harmless to the aquatic organisms but also not fit for drinking and other purposes. Overall, the trend of SS decreases from 2000 to 2014, reflecting an improvement in water quality. This positive impact can be explained by increased pollution control, improved waste water treatment, or improved practices of conservation.

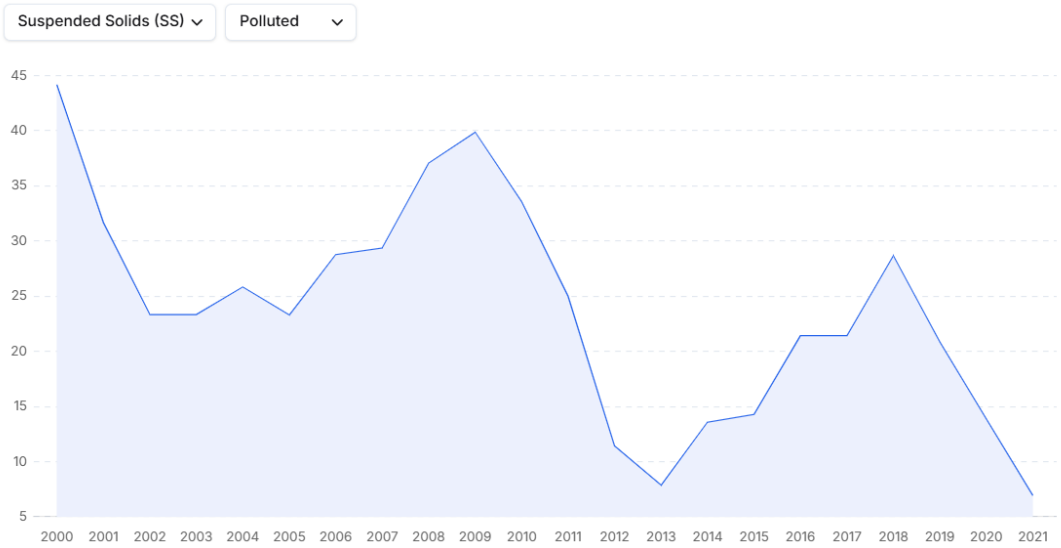


Figure 4.11: River Water Quality Stations Trend on *Suspended Solids (SS),* Sub-Index

All of these have slight increases in SS levels from 2014 until 2018, showing partial setbacks in keeping the levels of pollution low. This might also be because of increased industrial activities, construction, or changes in land use. From 2018 until 2021, the levels went down once more to reflect new impetus towards pollution control. Despite the progress, the slight increase in SS levels between 2014 and 2018 indicates that continuous and sustained efforts are essential to keep improving water quality and protecting aquatic ecosystems.

The analysis of water quality data for Biochemical Oxygen Demand (BOD5), Ammoniacal Nitrogen (NH3-N), and Suspended Solids (SS) from 2000 to 2021 reveals both challenges and progress in addressing water pollution. The trend of BOD5 shows an upward trend until 2014, reflecting the deteriorating pollution level, but then a sharp decline from 2014 to 2021, reflecting successful pollution control and eventual improvement in water quality. Similarly, NH3-N continued to increase until 2019, reflecting an increasingly serious nitrogen pollution problem; however, a sharp decrease in 2020 and 2021 reflected effective regulatory action and technological gains in treating wastewater.

The SS trend, in general, improved from the year 2000 up to 2014, then had a setback from 2014 to 2018, and continued to decline until 2021. This reflects the fact that, although the efforts put into the reduction of SS have generally been effective,

further vigilance and sustained actions are needed to maintain and further improve the water quality. Put together, these results again bring into focus the need for unceasing environmental interventions, regulatory measures, and improved technologies in pollution control for the health and safety of water resources in the long term.

4.3 Trend of River Water Quality Monitoring Stations

Figure 4.12, illustrates water quality trends from 2019 to 2023, categorizing water bodies as Clean, Slightly Polluted, or Polluted. Overall, there has been an improvement, with a notable increase in the percentage of clean water bodies and a decrease in polluted ones. In 2019, the percentage of clean water bodies was 62%, followed by 67% in 2020 and then 75% in 2021. Then it came down to 74% in 2022 and to 73% in 2023. Similarly, polluted water bodies decreased from 9% in 2019 to just 3% in 2021, but in subsequent years it slightly increased to 4%.

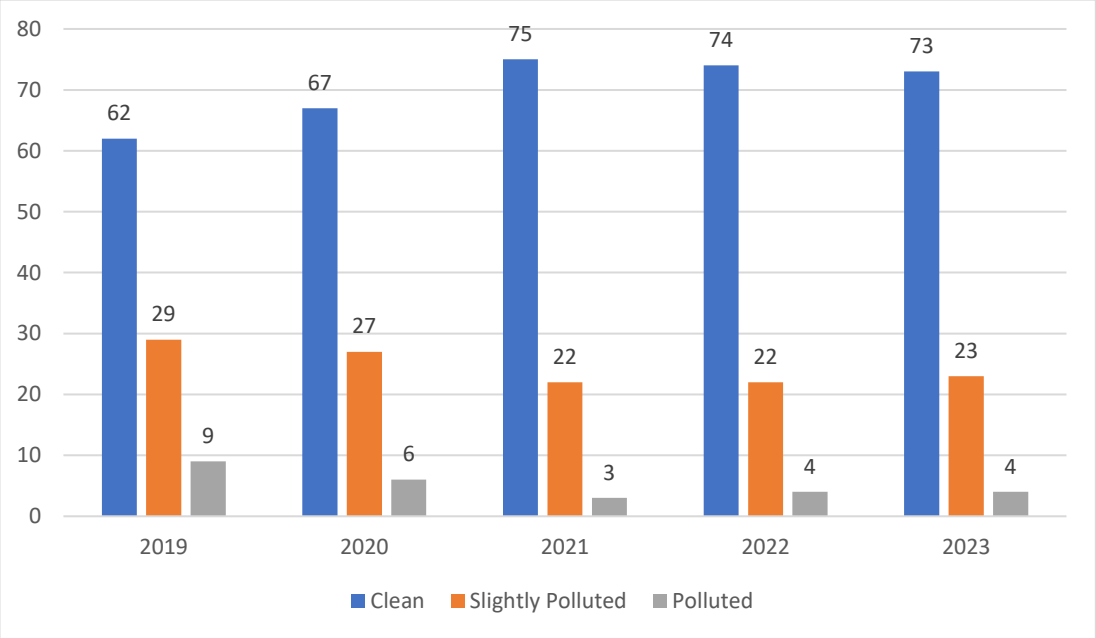


Figure 4.12: River Water Quality Stations Trend from 2019 until 2023

Although the trend is upward, there is a slight fluctuation every year. For example, water bodies that were slightly polluted decreased from 29% in 2019 to 22% in 2021 and rose slightly to 23% in 2023. Such fluctuations reveal that it is still

challenging to ensure improvement in water quality constantly and that further vigilance with effective measures against pollution is called for.

The water quality could have been greatly improved with pollution control legislation becoming stricter, methods of treating wastes becoming superior, and good practices by industries and farmers. This increased awareness and additional conservation efforts. However, the limitations in the data are that sources are not specified, and factors such as climate change and population growth are not considered; thus, while there is progress, more work needs to be done for sustained improvements to take place, with all influences on water quality tackled.

4.4 Trend of River Quality Monitoring on Sub-index

Figure 4.13, Trend of river water quality stations in BOD sub-index, 2019-2023, presents large variations on the trend of river water quality during the period under consideration. In 2019, only 16% of the stations representing the water quality of rivers were classified as clean, while a further 44% were slightly polluted and another 40% had been marked as polluted. In 2020, the percentage of clean stations reached as high as 34%, increased to 53% in slightly polluted stations, and drastically dropped to 13% for polluted stations. This trend was continued in 2021, with 78% of stations falling in the clean category, 13% being slightly polluted, and only 9% polluted-the biggest increase within these years.

This growth in 2021 was almost stable during the following years, considering only minor fluctuations. In fact, in 2022, all stations at 75%, slightly polluted at 14%, and 11% were polluted, showing a slight deterioration compared with the previous year. However, in 2023, the percentage of clean stations managed to rise to 78%, even if those slightly polluted went down to 12% and those polluted remained stable at 10%. This would go to imply that efforts at pollution control and quality management of water continue to pay off, though vigilance and further improvements must be made to keep such gains.

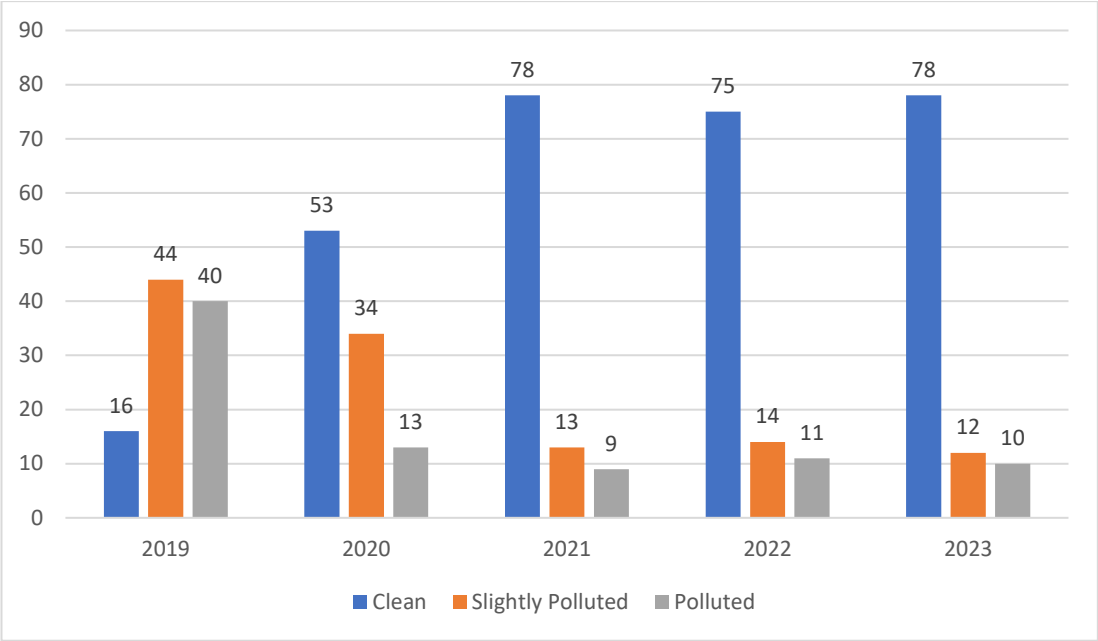


Figure 4.13: River Water Quality Station Trend on BOD Sub-index from 2019 until 2023

Figure 4.14 River water quality station trend on AN Sub-index from 2019 up to 2023, describes the share of river water quality stations in categories as clean, slightly polluted, and polluted within these years. In 2019, 34% of the stations were clean while 28% were slightly polluted and 38% polluted. In 2020, the figure dropped to 32% for clean stations, while Slightly Polluted stations went up to 30%, and those that were considered as Polluted stood at 38%. In the year 2021, the situation has considerably improved with a rise in clean stations to 51%, Slightly Polluted to 20%, and Polluted stations to 29%.

During 2022, it repeated with 49%, 24% more stations being Slightly Polluted, and the % coming down to 27% regarding the presence of a 'Polluted' level. For the year 2023, the numbers were 51%, those in slightly polluted decreased to 23%, while polluted went further to 26%. These changes indicate a general improvement in the water quality of the province, ups and downs in the percentage variations of Slightly Polluted and Polluted stations, while the number of clean stations is growing across the years. The improvement points to effective environmental measures and better efforts on pollution control, though there is a need for continued attention in this respect.

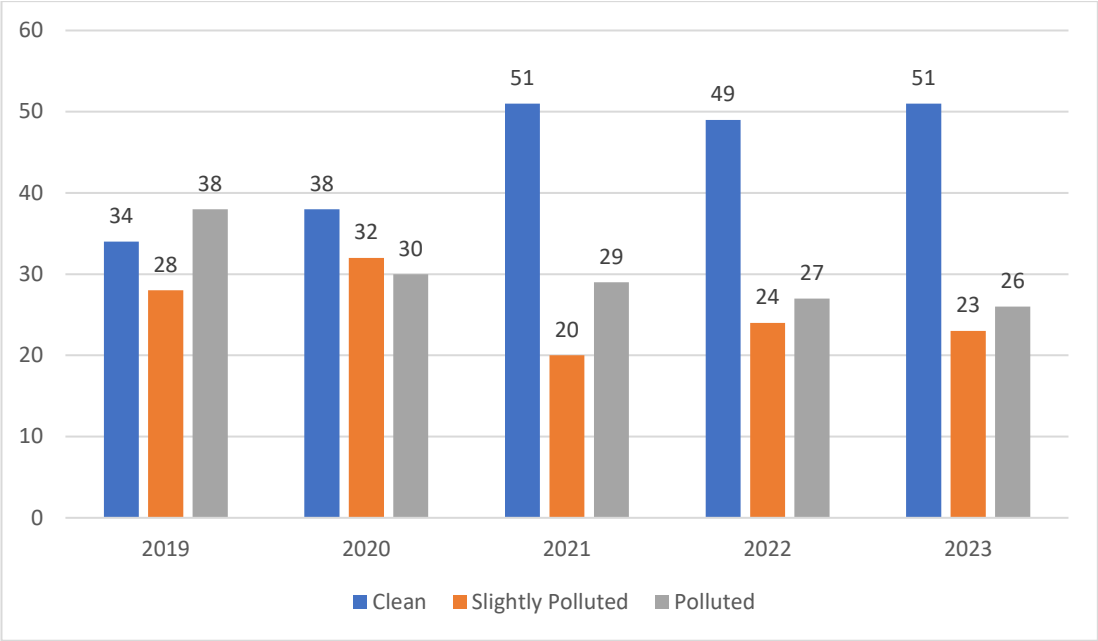


Figure 4.14: River Water Quality Station Trend on AN Sub-index from 2019 until 2023

Figure 4.15 Trend of SS Sub-index for River Water Quality Station from 2019 up to 2023. The percentage share of river water quality stations classified as clean, slightly polluted and polluted are given for the mentioned years. For example, in the year 2019, the percentage share of the classified station as a clean category was 62%, slightly polluted was 12% while polluted ones reached 26%. The following year, 2020, saw a slight improvement with 63% clean stations, 11% slightly polluted stations, and 26% polluted stations remaining constant. There was a significant improvement in 2021, with 77% clean stations, a reduction to 9% for slightly polluted stations, and a further drop in polluted stations to 14%.

In 2022, this percentage decreased to 74%, while those considered slightly polluted remained at 8%, and those considered polluted increased to 18%. In 2023, this trend continued at 74% for clean stations, 8% for slightly polluted, and polluted ones at 18% stability. This shows the overall development of the water quality in these five years, with distinct improvement between 2019 and 2021. Some minor fluctuations afterward underline that it is of utmost importance to continue maintaining and further enhancing the quality by systematic monitoring and adequate measures for pollution control.

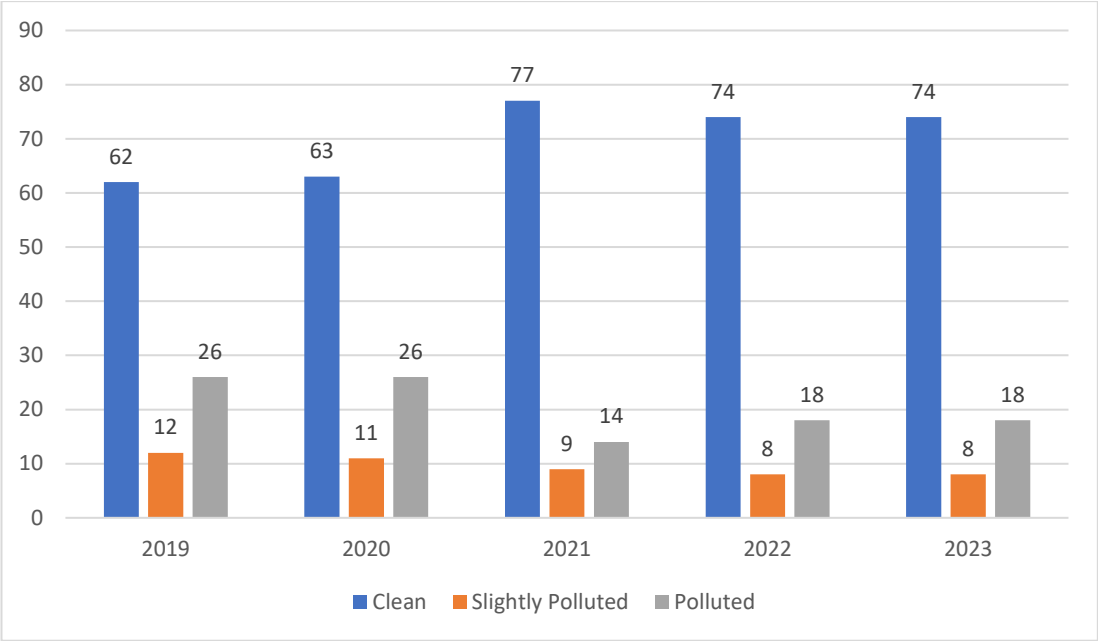


Figure 4.15: River Water Quality Station Trend on SS Sub-index from 2019 until 2023

River water quality, on the other hand, for 2019 and projected to 2023, presented some improvements and challenges on BOD, Ammoniacal Nitrogen, and Suspended Solids. In contrast, BOD had a greater improvement of higher clean and lesser polluted stations. This change became more visible in the year 2021 and after that. Correspondingly, its AN sub-index reported progress and, in general, an increase of a high rate in clean stations for 2021, maintaining positive tendencies with some oscillation during the following years.

The SS sub-index also improved, with more stations becoming clearer and less polluted, although there was a moderate fluctuation year by year. In general, these trends reflect the positive direction in water quality and effective measures for pollution control. However, this requires further efforts and adaptive strategies to maintain the gained pace and to overcome various challenges in water quality management.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Summary

This research aimed to develop and evaluate advanced machine learning models for predicting water pollution levels in Malaysian river basins. The study used a holistic dataset of historical water quality data, meteorological factors, and other relevant parameters obtained from DOSM. A critical part of the research work was the collection and preprocessing of water quality data from DOSM with a great deal of caution, considering accuracy, consistency, and suitability for machine learning model training. This has used some heavy data cleaning and preprocessing methodology for handling missing values, outliers, and inconsistency in the dataset.

A range of machine learning algorithms was studied and assessed on their capabilities with respect to effective predictions of pollution levels. Among them were well-known powerful methods like Random Forest and Long Short-Term Memory-LSTM networks. It involved extensive tuning of hyperparameters to optimize each model's performance with the dual goal of maximizing its predictive accuracy and generalizability. Performance for these developed models is investigated rigorously with a suite of relevant evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. This was to ascertain the most appropriate and reliable model that could deliver real-time or near-real-time pollution level prediction with higher accuracy.

The study identified many improvements within the water quality of Malaysian river basins, specifically for the following key indicators: Biochemical Oxygen Demand, BOD5, and Suspended Solids, SS. The upward trend of this series may reflect the effectiveness of continuous environmental interventions and policies. Some of the challenges faced in the development and deployment of such predictive models were underlined in the research. This would include good quality and consistent

availability of data from multiple sources, interpretability of model output mechanisms, and smooth integration with current hydrological models to achieve complete and accurate predictions using machine learning models. Overall, the current study provides an overview of the role that machine learning could play in the further development of the potentials for water quality monitoring and management within Malaysia. These findings will be useful as guidelines for effective, data-driven strategy development aimed at mitigating pollution to protect water resources across the nation.

5.2 Future Works

The findings from this research give a very strong foundation for further studies in this field. Some of the possible promising directions that could be explored in order to further enhance the understanding and, therefore, prediction of pollution in river basins include the exploration of advanced machine-learning techniques. Deep learning architectures, such as convolutional neural networks and LSTM networks, can be employed in order to capture complex temporal and spatial patterns in pollution dynamics. Furthermore, hybrid models that marry strengths of physical and statistical approaches can result in better predictability with more insight into the process.

Integrating diverse data sources would largely help in improving the accuracy and completeness of the predictions regarding pollution. The use of IoT sensors, satellite imagery, and real-time data streams can give a better view of the patterns of pollution that could enable timely, more accurate interventions. Furthermore, explainable AI-XAI needs to be developed. The XAI models can make transparent and interpretable predictions that shall help stakeholders understand how the model made decisions and engender trust in the model outputs. This aspect of transparency acts as a key driver for effective communication and successful implementation of model-driven interventions.

Finally, broadening the scope of research to cover global modeling of river basins can make useful inferences on the pattern and management of pollution in different geographical scopes. This would, in turn, demand global models that take into consideration the river basin characteristics for achieving an appropriate and

effective approach to managing pollution globally. Besides, a study of socioeconomic impacts of pollution should be assessed in terms of human health, agriculture, tourism, among other sectors, which will ultimately help in formulating effective policies and mitigation strategies with consideration of multifaceted consequences of river basin pollution. These are the future research directions which will help us to come up with more accurate, reliable, and interpretable predictive models for river basin pollution. This will ensure better environmental management, sustainable water resource management, and a healthy planet for the future generation.

REFERENCES

Agarwal, A., & Singh, S. (2018). Seasonal variations in water pollution and its management. *Environmental Monitoring and Assessment*, 190(3), 174. https://doi.org/10.1007/s10661-018-6531-5

Aggarwal, C. C. (2015). *Outlier analysis* (2nd ed.). Springer.

Ahmad, T., Khan, S., & Malik, R. (2021). Predictive analytics in environmental impact assessments. *Environmental Modeling & Assessment*, 26(4), 345-358.

Ahmed, S., Zhai, L., & Lee, J. (2020). Addressing the challenges of machine learning in environmental monitoring: A review of prediction accuracy and interpretability. *Environmental Modelling & Software*, 129, 104700. https://doi.org/10.1016/j.envsoft.2020.104700

Almasri, M., et al. (2020). Addressing data challenges in environmental modeling. *Journal of Hydrology*, 590, 125-134.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Brown, J., et al. (2020). Best practices for training and testing in predictive modeling. *Environmental Research*, 186, 109-118.

Chatfield, C. (2003). *The analysis of time series: An introduction* (6th ed.). CRC Press.

Chapra, S. C. (2008). *Surface water-quality modeling*. Waveland Press.

Chen, L., et al. (2021). LSTM applications in water pollution prediction. *Water Research*, 201, 117-130.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Gao, Z., et al. (2019). Feature engineering for river basin modeling. *Science of the Total Environment*, 651, 112-123.

Gao, X., Zhang, L., & Li, S. (2021). Real-time environmental monitoring and prediction using IoT-based sensors: Applications in pollution management. *Environmental Monitoring and Assessment*, 193(2), 75. https://doi.org/10.1007/s10661-021-08796-5

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

Gholami, V., Abdi, R., & Kamali, B. (2021). Application of machine learning techniques for water quality prediction: A review. *Environmental Science and Pollution Research*, 28(1), 468-488. https://doi.org/10.1007/s11356-020-10447-y

Goodchild, M. F., Parks, B. O., & Steyaert, L. T. (1992). *Environmental modeling with GIS*. Oxford University Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680. https://papers.nips.cc/paper/5423-generative-adversarial-nets

Gupta, A., Kumar, M., & Das, S. (2020). Mapping pollution hotspots in the Ganges basin using machine learning: A clustering approach. *Environmental Science and Pollution Research*, 27(19), 24316-24328. https://doi.org/10.1007/s11356-020-08763-x

Gupta, R., et al. (2020). Pollution hotspot mapping using machine learning. *Journal of Environmental Management*, 262, 110284.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hassan, M., Lee, J., & Shamsuddin, A. (2019). Predictive analytics in occupational safety: A framework for risk reduction. *Journal of Safety Research*, 70, 89-98. https://doi.org/10.1016/j.jsr.2019.02.001

Huang, C., Shen, Y., & Chang, H. (2020). Predicting river water quality using ensemble machine learning models. *Journal of Hydrology*, 584, 124660. https://doi.org/10.1016/j.jhydrol.2020.124660

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

James, K., & Dunn, R. (2020). The role of machine learning in enhancing occupational health and safety practices. *Safety Science*, 124, 104601. https://doi.org/10.1016/j.ssci.2020.104601

Jones, D., & Taylor, A. (2019). Machine learning in policy-making for water resources. *Water Policy*, 21(3), 456-467.

Khan, A., et al. (2020). Comprehensive water quality monitoring frameworks. *Environmental Monitoring and Assessment*, 192, 245.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kumar, A., & Singh, V. (2020). Application of predictive models in river basin management for pollution hotspot identification. *Water Research*, 180, 115861. https://doi.org/10.1016/j.watres.2020.115861

Kumar, P., & Singh, R. (2019). Long-term predictive models for water quality. *Ecological Modelling*, 404, 89-98.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, X., Zhang, S., & Chen, J. (2019). Source attribution of water pollution using predictive models: A case study in a river basin. *Environmental Science & Technology*, 53(24), 14612-14620. https://doi.org/10.1021/acs.est.9b04478

Li, X., et al. (2021). Gradient Boosting for pollution prediction. *Science of the Total Environment*, 752, 141759.

Li, X., Zhang, Y., Wang, H., & Chen, Y. (2023). A hybrid machine learning model for water quality prediction in a complex river basin. *Journal of Hydrology*, 625, 129300.

Liu, Y., et al. (2022a). Real-time analytics in predictive modeling. *Journal of Water Resources Planning and Management*, 148(2), 04022014.

Liu, Y., Zhang, W., & Li, X. (2022b). Advancing predictive models for environmental monitoring: The role of IoT sensors and real-time analytics. *Environmental Research Letters*, 17(7), 074006. https://doi.org/10.1088/1748-9326/ac72b9

Mason, J., Nguyen, M., & Tan, S. (2021). Data-driven safety decision-making: Insights from predictive models. *Journal of Hazardous Materials*, 417, 125753. https://doi.org/10.1016/j.jhazmat.2021.125753

Miller, K., et al. (2020). Short-term water quality predictions. *Journal of Environmental Sciences*, 94, 35-45.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. Wiley.

Novotny, V. (2003). *Water quality: Diffuse pollution and watershed management*. John Wiley & Sons.

Patel, H., Prajapati, J., & Solanki, D. (2020). Comparative study of error metrics in predictive modeling. *International Journal of Data Science and Analytics, 6*(4), 273-285. https://doi.org/10.1007/s41060-020-00221-1

Patel, R., et al. (2020). Evaluation metrics in predictive modeling. *Environmental Informatics*, 42, 77-92.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

Rahman, M., & Akhtar, S. (2020). Data preprocessing in environmental data science. *Environmental Data Science*, 3(1), 12-25.

Schwalm, C. R., He, L., & Mao, J. (2020). Remote sensing applications for environmental monitoring: Implications for water quality and pollution management. *Environmental Monitoring and Assessment*, 192(4), 260. https://doi.org/10.1007/s10661-020-8067-4

Seitzinger, S. P., Harrison, J. A., Bohlke, J. K., Bouwman, A. F., & Caraco, N. F. (2010). Denitrification across landscapes and waterscapes: A synthesis. *Ecological Applications*, 20(4), 1536-1552. https://doi.org/10.1890/09-0151.1

Shao, S., Yang, X., & Zhang, J. (2019). Community involvement and the effectiveness of water pollution control policies: Evidence from a machine learning approach. *Environmental Research Letters*, 14(3), 034012. https://doi.org/10.1088/1748-9326/aafc5f

Singh, V., et al. (2021). GIS-integrated machine learning for river analysis. *Hydrological Processes*, 35(5), e14123.

Smith, J., et al. (2018). Pollution source identification methods. *Journal of Cleaner Production*, 174, 1015-1027.

Tao, J., Zhang, J., & Liu, S. (2021). Integrating remote sensing data with machine learning for pollution prediction: A case study of water quality monitoring. *Environmental Science & Technology*, 55(15), 10318-10328. https://doi.org/10.1021/acs.est.1c03593

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., ... & Kaiser, Ł. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need

Vörösmarty, C. J., McIntyre, P. B., & Gessner, M. O. (2010). Global threats to human water security and river biodiversity. *Nature*, 467(7315), 555-561. https://doi.org/10.1038/nature09440

Wang, X., et al. (2020). Clustering-based water quality analysis. *Journal of Environmental Management*, 261, 110234.

Ward, R. C., & Robinson, M. (2000). *Principles of hydrology* (4th ed.). McGraw-Hill.

Xu, Y., Li, J., & Zhang, L. (2020). IoT-based pollution monitoring and prediction in water bodies: A review of technologies and applications. *Environmental Science & Technology*, 54(6), 3401-3413. https://doi.org/10.1021/acs.est.9b06968

Yang, X., Wang, Q., & Zhang, Z. (2018). Machine learning approaches for water quality prediction in river systems: A comparative study. Environmental Monitoring and Assessment, 190(12), 707. https://doi.org/10.1007/s10661-018-7096-y

Zhang, H., et al. (2019). Multi-point water quality monitoring. *Water Science and Technology*, 79(4), 742-750.

Zhang, T., et al. (2021a). Cross-validation techniques for model robustness. *Journal of Hydrology*, 595, 125-135.

Zhang, Y., Bengio, S., & Hardt, M. (2021b). *Understanding generalization in machine learning: A review*. Journal of Machine Learning Research, 22, 1-50.

Zhang, Y., Li, X., Wang, H., & Chen, Y. (2022). Deep learning for algal bloom prediction in lakes: A review. *Water Research*, 218, 118467.

Zhou, S., Yu, L., & Zhang, Y. (2021). Hybrid models for environmental prediction: Balancing accuracy and simplicity. *Environmental Science & Technology*, 55(6), 3184-3192. https://doi.org/10.1021/acs.est.0c07354

Zhang, Z., Wang, X., & Liu, H. (2021). The role of machine learning in predicting river pollution and its potential applications. *Environmental Monitoring and Assessment*, 193(5), 296. https://doi.org/10.1007/s10661-021-09012-8