3. PROBLEM MODELLING

The main focus of this research is to propose an evident model to predict water quality index for river water mainly for the Bhavani River and the Bharathapuzha River. The research problem of predicting water quality is formulated as a regression task and a suitable solution is proposed using deep learning architectures and transfer learning approaches. This chapter portrays the approach of problem modelling that facilitates the objectives. The creation of four independent datasets consisting of the water quality features is also described in this chapter. The WQI prediction models developed using deep learning techniques and transfer learning approaches in four phases are elucidated in detail in this chapter.

3.1. OVERALL FRAMEWORK OF THE WQI PREDICTION MODEL

The methodology designed to generate the WQI prediction model comprises four modules. They are (i) data acquisition (ii) exploratory data analysis, data preprocessing and dataset preparation (iii) building the WQI prediction model and (iv) validation and model evaluation.

A five-year time series data comprising 26 physiochemical water quality parameters observed from eleven sampling stations located along the Bhavani River, for the period 1st January 2016 to 31st December 2020 are collected in real-time. Also, data are collected from three sampling stations of the Bharathapuzha River for the period 1st January 2020 to 31st December 2021. Physical parameters such as temperature, total suspended solids, turbidity, fixed dissolved solids, conductivity, and total dissolved solids, as well as chemical parameters such as pH, ammonia, alkalinity, chloride, potassium, sulphate, nitrogen, fluoride, hardness, dissolved oxygen, biological oxygen demand, and chemical oxygen demand, are taken for study. Total coliform and faecal coliform are also measured as biological indicators of water quality. Seasonal variations play a critical role in assessing water quality, as changes in weather and atmospheric conditions are reflected in seasonal parameters. The seasonal parameters include temperature, dew point, humidity, sea level pressure, precipitation, precipitation amount, wind speed, wind direction, cloud cover, and visibility. These seasonal data are collected from visual crossing sites for the same period and locations. Spatial parameters such as latitude, longitude, station ID and temporal parameter Date are also taken into consideration for the study. The WQI is calculated according to the Bureau of Indian Standards for drinking water specification and augmented with time series data.

The data collected on river water quality is subjected to exploratory data analysis to comprehend the properties of the data and evaluate the significance of each parameter in generating the water quality index. Various analysis methods such as heatmap analysis, boxplot analysis, histogram analysis and pair plot analysis are employed. The preprocessing tasks such as handling missing values, removal of outliers, and data normalization is done. Min-max normalization is used to normalize the observations. Feature selection technique namely Select K Best is applied and the most relevant independent variables that contribute significantly to predicting the WQI are identified. Three datasets namely WQI-PCA, WQI-SA and WQI-BP are developed for building deep learning-based WQI prediction models.

Here, the problem of WQI prediction is modelled as a regression task and an appropriate solution is obtained using deep learning. Regression analysis is a statistical technique commonly used for data fitting and prediction purposes. The mathematical model of regression describes the relationship between a dependent variable (Y) and a set of independent variables (Xi). The model is expressed as $Y = \beta_0 + \beta_1*X_1 + \beta_2*X_2 + ... + \beta_n*X_n + \varepsilon$, where Y represents the dependent variable. The intercept (β_0) is the constant term, and β_1, β_2, ..., β_n are the coefficients associated with the independent variables X_1, X_2, ..., X_n, respectively. The independent variables X_1, X_2, ..., X_n are used to explain or predict the value of the dependent variable. The error term (ε) accounts for unexplained variability or noise in the model. The goal of regression analysis is to estimate the coefficients (β_0, β_1, β_2, ..., β_n) that provide the best fit to the data, minimizing the difference between the observed values of Y and the predicted values based on the independent variables.

Data fitting for regression is modelled by defining the different categories of water quality parameters as independent variables (X_i) and WQI as the target variable (Y). The deep learning and transfer learning prediction models are built using various architectures, including RNN, LSTM, GRU, and TFT. The research is carried out in four phases. In the first phase, deep learning architectures such as RNN, LSTM and GRU are utilized as these architectures are significant in training sequence data and accurate WQI prediction models are developed. Next, a modern architecture called Temporal Fusion Transformer is employed and efficient WQI prediction models are built by training the same datasets. In the third phase, the homogenous transfer learning technique is adopted for building the enhanced WQI prediction model. In the fourth phase, the

heterogeneous transfer learning approach is adopted for constructing a hybrid WQI prediction model.

The performance of the WQI prediction models is evaluated with various metrics such as mean absolute error, mean squared error, root mean squared error and the R2 score and the result analysis is done. The proposed framework of the water quality index prediction model is shown in Fig.3.1 and the various tasks of the proposed methodology are explained in detail in the following sections.

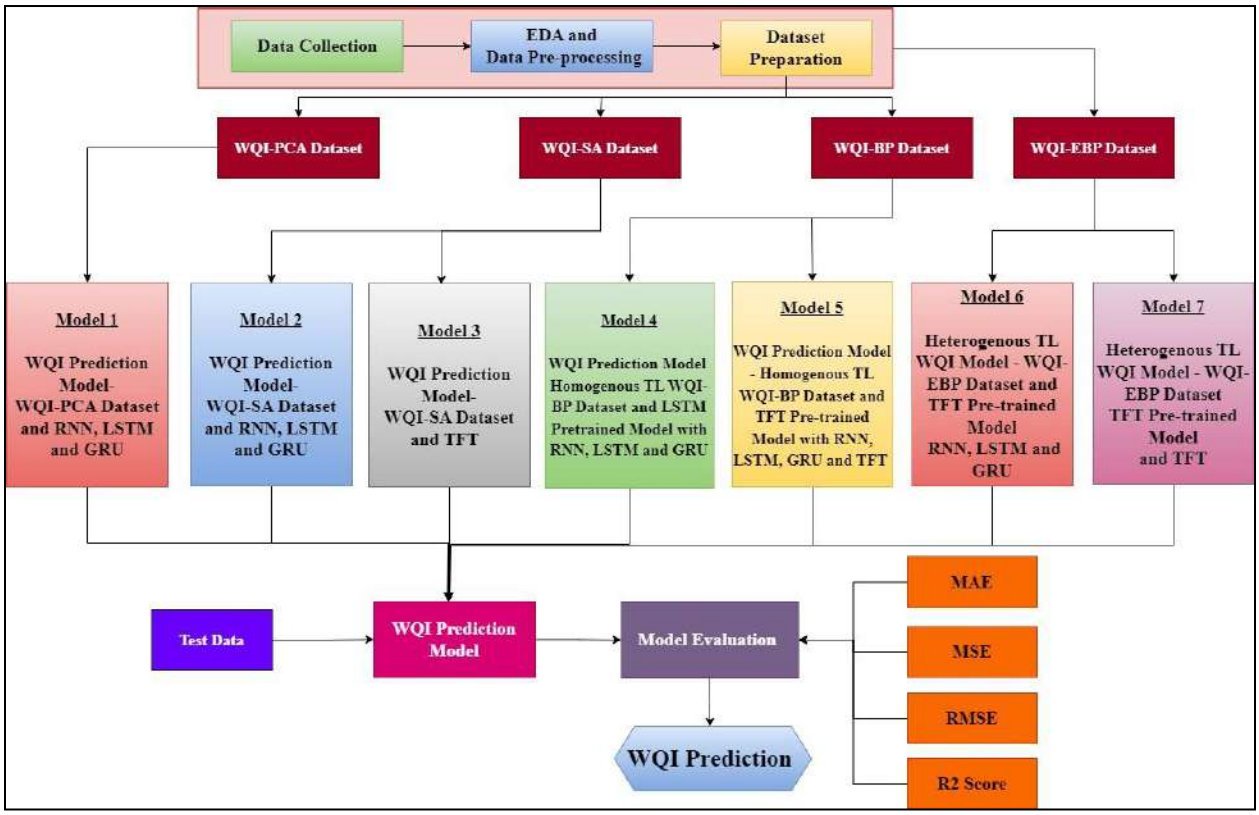

Fig. 3.1 Overall Framework of Water Quality Index Prediction Model

3.2. DATA COLLECTION

Data collection is the process of gathering and storing information for analysis and decision-making purposes. The data collection for this research involves gathering crucial information from two major rivers, namely the Bhavani River and Bharathapuzha River. The observation for water quality parameters monitored during the period January 2016 to December 2020 across 11 monitoring stations of Bhavani River, and the observations monitored during the period January 2019 to December 2020 across 3 sampling stations of Bharathapuzha River are

taken for study. The extensive collection of data from multiple monitoring stations provides a comprehensive understanding of the water quality dynamics in these rivers over the specified time periods.

The purpose of this research is to develop reliable models for predicting the WQI of river water based on water quality parameters, such as physical, chemical, biological and seasonal. When the parameter values associated with water quality change, it can have significant effects on the overall condition of the water. Parameters such as pH, dissolved oxygen levels, temperature, turbidity, nutrient concentrations, and pollutant levels can greatly impact water quality. For instance, an increase in nutrient concentrations, such as nitrogen and phosphorus, leads to eutrophication and harmful algal blooms, resulting in decreased water quality. Changes in pH levels affect the acidity or alkalinity of the water, influencing the health of aquatic organisms. Similarly, alterations in temperature impact metabolic rates and habitat suitability for various species. Therefore, understanding and monitoring parameter values are crucial in assessing and maintaining water quality for various ecosystems and human needs.

BHAVANI RIVER

The Bhavani River is a significant water body located in the Tamil Nadu and Kerala states of India. It originates from the Silent Valley National Park in Kerala and flows through the Coimbatore district of Tamil Nadu before joining the Kaveri River. The river has a total length of about 217 kilometres and serves as a major source of irrigation and drinking water for the surrounding communities. Despite its vital importance, the Bhavani River has been facing significant pollution levels due to industrial and domestic waste discharges, deforestation, and other human activities. The high levels of pollution have led to the deterioration of the river's water quality and have negatively impacted the health of the local ecosystem, including the aquatic life and the surrounding vegetation. The river is also prone to flash floods during heavy rains, which further exacerbate pollution levels and threaten the safety of the local population. The government and other stakeholders are taking various measures to address the pollution levels and restore the health of the Bhavani River. However, significant challenges remain, and ongoing efforts are necessary to ensure the sustainable management and conservation of this vital water resource.

There are eleven monitoring stations situated in Bhavani River. These monitoring stations are located at Kottathara, Thavalam, Chalayur, Karathur, Cheerakuzhi, Elachivazhi,

Badrakaliamman kovil, Sirumugai, Bhavanisagar, Bhavani, Sathyamangalam. The water flow of the Bhavani River is depicted in Fig.3.2.

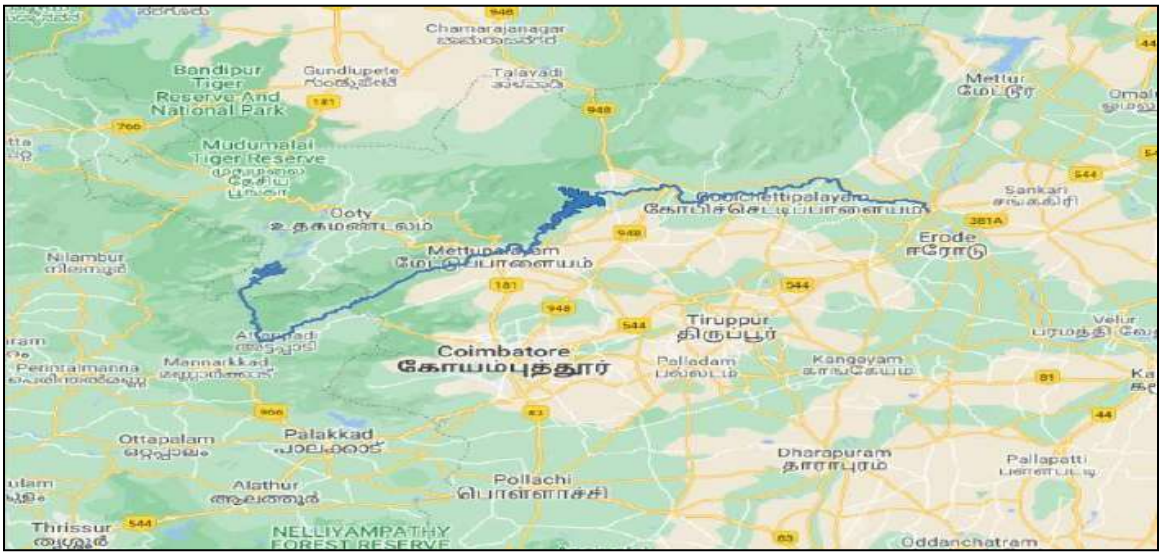

Fig. 3.2 The Flow of the Bhavani River

BHARATHAPUZHA RIVER

Bharathapuzha River is a significant river system located in the southern Indian state of Kerala. It is also known as the river Nila and is the second-longest river in Kerala, spanning a length of approximately 209 kilometres. The river originates from the Anaimalai Hills in Tamil Nadu and flows through several districts of Kerala, including Palakkad, Thrissur, and Malappuram, before draining into the Arabian Sea. The Bharathapuzha River has immense ecological and socio-economic significance to the region, supporting a range of aquatic flora and fauna and providing irrigation, drinking water, and livelihoods for millions of people. The river is facing significant pollution and degradation due to various anthropogenic activities, such as agricultural run-off, industrial effluent discharge, and urbanization. The river water quality has become a matter of great concern for water managers, and other stakeholders. Various studies have been conducted to assess and monitor the water quality of the Bharathapuzha River, and it is imperative to continue such efforts to ensure the sustainable development and conservation of this vital river system. There are three different sampling stations situated across the Bharathapuzha River. These sampling stations are located at Kuttipuram, Pattambi and Korayar Kanjikode. The water flow of Bharathapuzha river is shown in Fig.3.3.

Fig. 3.3 The Flow of the Bharathapuzha River

It is essential to continue monitoring and assessing the water quality of these rivers to ensure sustainable development and the well-being of both the human population and the aquatic ecosystem. Accurate and reliable data on various water quality parameters, like pH, dissolved oxygen, turbidity, and nutrients, among others, are essential for developing reliable models and methods for predicting the WQI of these rivers. Moreover, collecting data over time helps to identify trends in water quality and understand the sources of pollution and their impacts.

WATER QUALITY PARAMETERS

Accurate prediction of the WQI necessitates the availability of dependable and inclusive data encompassing multiple parameters. Collecting accurate and reliable data on these parameters is very important for this study. Various parameters that are required to build the prediction model or WQI prediction are classified into four categories: physical, chemical, biological and seasonal parameters.

Physicochemical parameters are properties of water that relate to its physical and chemical characteristics. These parameters play a crucial role in determining the water quality of a particular source. The physical parameters include temperature, conductivity, turbidity, total suspended solids, and fixed dissolved solids. The chemical parameters of water quality such as pH, ammonia, alkalinity, chloride, potassium, sulphate, nitrogen, fluoride, hardness, dissolved oxygen, biological oxygen demand, and chemical oxygen demand. The biological water quality indicators are total coliform and faecal coliform. Seasonal parameters refer to variables that exhibit regular patterns or fluctuations based on the time of year, allowing for effective analysis and prediction of seasonal

trends. The seasonal parameters include dew, humidity, sea level pressure, precipitation, precip over, wind speed, wind direction, cloud cover, and visibility.

Physical Parameters

Physical parameters such as temperature, conductivity, turbidity, total suspended solids, and fixed dissolved solids, significantly affect the chemical and biological properties of water and they are described below.

Temperature

Temperature plays a crucial role in predicting the WQI, serving as a measure of water's hotness or coldness and impacting its physical, chemical, and biological properties. High water temperatures promote biological activity, leading to algae and bacteria growth that deplete oxygen levels and affect aquatic organisms' survival. Conversely, low temperatures restrict biological activity and reduce dissolved oxygen. Temperature also influences the solubility of minerals and chemicals, altering water chemistry. Measuring temperature at various depths helps assess thermal stratification, crucial for nutrient and oxygen distribution. Understanding temperature patterns is vital for predicting and managing water quality. The WQI calculates by determining the deviation from the standard value, which is 25°C according to the Bureau of Indian Standards (BIS).

Turbidity

Turbidity is an essential parameter used to predict the WQI and assess the clarity of water. Elevated turbidity levels indicate a cloudy or hazy appearance due to suspended particles originating from natural sources or human activities. These particles can include sediment, organic matter, and runoff from construction or agriculture. Turbidity influences light penetration, impacting aquatic plant growth and the survival of organisms. The WQI considers a minimum turbidity value of 0 Nephelometric Turbidity Units (NTU) and a maximum value of 1000 NTU. Turbidity is measured with a turbidimeter, and the WQI is calculated by determining the deviation from the standard value. The BIS has established a surface water turbidity standard of 5 NTU, with the deviation calculated by subtracting the measured turbidity from 5 NTU and dividing by 1. BIS also defines a maximum permissible limit of 25 NTU for surface water turbidity, beyond which it is deemed unsuitable for human consumption.

Conductivity

Conductivity is an essential parameter used in the prediction of the Water Quality Index and measures the ability of water to conduct electrical current, reflecting the concentration of dissolved salts and minerals. High conductivity values indicate elevated concentrations of dissolved minerals, while low values suggest lower concentrations. The WQI considers a minimum conductivity value of 0 micro siemens per centimetre (uS/cm) and a maximum value of 1000 uS/cm. Measuring conductivity is done using a conductivity meter, and the WQI calculation involves determining the deviation from the standard value. The BIS has established a surface water conductivity standard of 75 uS/cm, with the deviation calculated by subtracting the measured conductivity from 75 uS/cm and dividing by 25. BIS has also specified a maximum permissible limit of 2250 uS/cm for surface water conductivity, beyond which it is considered unsuitable for human consumption.

Total Suspended Solids

Total Suspended Solids (TSS) is an important parameter used in the prediction of the Water Quality Index. TSS is the amount of solid material that is suspended in water and includes particles such as silt, clay, and organic matter. High TSS levels can reduce water clarity, interfere with aquatic life, and affect water treatment processes. The minimum value considered for WQI prediction is 0 milligrams per litre (mg/L), while the maximum value is 500 mg/L. TSS is measured using a filter and gravimetric analysis, and the WQI is calculated by determining the deviation of the measured TSS from the standard value. The BIS has set a standard value of 10 mg/L for TSS in surface water, and the deviation from the standard value is calculated by subtracting the measured TSS from 10 mg/L and then dividing by 2. The BIS has also specified a maximum permissible limit of 100 mg/L for TSS in surface water, beyond which it can be considered unsuitable for human consumption.

Total Dissolved Solids

Total Dissolved Solids (TDS) is an important parameter used in the prediction of the WQI. TDS is the amount of inorganic and organic materials that are dissolved in water and includes minerals, salts, and organic compounds. High TDS levels can affect water taste, interfere with aquatic life, and increase the risk of scaling and corrosion in pipes and equipment. The minimum value considered for WQI prediction is 0 milligrams per litre (mg/L), while the maximum value is

2000 mg/L. TDS is measured using a gravimetric analysis or conductivity meter, and the WQI is calculated by determining the deviation of the measured TDS from the standard value. The BIS has also specified a maximum permissible limit of 2000 mg/L for TDS in surface water, beyond which it is considered unsuitable for human consumption.

Fixed Dissolved Solids

Fixed Dissolved Solids (FDS) is a significant parameter used in the prediction of the WQI. FDS refers to the inorganic materials remaining in solution after evaporation, including calcium, magnesium, and chloride ions. High FDS levels impact water taste and can lead to scaling and corrosion in pipes and equipment. The WQI considers a minimum value of 0 mg/L and a maximum value of 1500 mg/L for FDS. FDS is measured using gravimetric analysis or a conductivity meter, and the WQI calculates the deviation from the standard value. BIS has also specified a maximum permissible limit of 1500 mg/L for FDS in surface water, beyond which it is unsuitable for human consumption.

Chemical Parameters

Chemical river water quality parameters refer to the various chemical components present in water that affect its suitability for different uses. These parameters are essential for understanding the overall health of a river and its potential impact on the environment and human health. Some of the key chemical parameters used in river water quality assessment include pH, ammonia, alkalinity, chloride, potassium, sulphate, nitrogen, fluoride, hardness, dissolved oxygen, biological oxygen demand, and chemical oxygen demand and these are described below.

pH

pH is a critical chemical parameter used in river water quality assessment. It is a measure of the acidity or basicity of water and has a significant impact on the chemical and biological processes in water bodies. The minimum pH value considered for water quality index prediction is 0, which represents highly acidic water, while the maximum value is 14, which represents highly alkaline water. The pH is measured on a scale of 0 to 14 using a pH meter or pH indicator strips. In the context of river water quality index calculation, the pH value is compared to a standard value, which is typically set at 7.0. The deviation from the standard value is then calculated by subtracting the measured pH value from 7.0. The BIS has specified a permissible limit for pH in

river water between 6.5 and 8.5. pH levels outside this range can have adverse effects on aquatic life and indicate potential pollution sources.

Ammonia

Ammonia is one of the chemical parameters used for river water quality index prediction. It is a colourless, pungent gas that dissolves readily in water to form ammonium ions. Ammonia levels in river water increase due to various human activities such as agriculture, sewage treatment, and industrial discharge. The minimum and maximum permissible limits for ammonia in river water according to the BIS are 0.5 and 2.0 mg/L, respectively. The concentration of ammonia is measured using various methods, including colourimetry and ion-selective electrodes. The deviation from the standard value of ammonia is calculated by subtracting the measured value from the permissible limit of 1.25 mg/L. Therefore, regular monitoring of ammonia levels in river water is crucial to maintain the quality and health of the water body.

Alkalinity

Alkalinity is an important chemical parameter used in the prediction of river WQI. It measures the water's capacity to neutralize acids and maintain a stable pH level. Alkalinity is influenced by the presence of dissolved bicarbonates, carbonates, and hydroxides in the water. These compounds are naturally occurring in river water and can be affected by human activities such as agriculture and industrial activities. The minimum value for alkalinity in river water is 20 mg/L, while the maximum value is 600 mg/L. Alkalinity is calculated by titrating a water sample with acid until the pH drops to a specific level. The amount of acid required to reach this level is then used to determine the alkalinity. The standard value for alkalinity in surface water is 200 mg/L, according to the BIS. A higher alkalinity value indicates better water quality, as it can buffer against pH changes and maintain a stable aquatic environment.

Chloride

Chloride is a chemical parameter used for the prediction of the river water quality index. It is an anion commonly found in natural water sources and is an important indicator of water salinity. High chloride levels indicate contamination from sources such as road salts, industrial waste, and agricultural runoff. Chloride is measured in milligrams per litre (mg/L) and the minimum and maximum permissible levels for river water, according to the BIS, are 250 mg/L and 1000 mg/L, respectively. The deviation of chloride levels from the standard value of 250 mg/L is used in the

calculation of WQI. The WQI score decreases with increasing deviation from the standard value. Chloride affects the taste of water and high levels can also corrode pipes and damage aquatic habitats. Understanding chloride levels in river water is important for maintaining water quality and ensuring safe usage.

Potassium

Potassium is one of the essential parameters used in the prediction of river water quality index. It is a crucial element in the growth and development of aquatic plants and is an indicator of the presence of nutrients in the water. The minimum value of potassium that is considered acceptable for river water is 2 mg/L, while the maximum value should not exceed 12 mg/L. The calculation of potassium levels in river water is typically done through laboratory analysis using spectrophotometry or atomic absorption spectrometry. The BIS has set a standard value of 6 mg/L for potassium in river water, which serves as a benchmark for water quality assessment.

Sulphate

Sulphate is an important parameter used in the prediction of river water quality index. It is a naturally occurring compound found in rocks and soils, and can also be present in water due to human activities such as mining and industrial discharges. The acceptable range for sulphate levels in river water is typically between 200-400 mg/L, with a maximum limit of 1000 mg/L to prevent adverse effects on aquatic life. The calculation of sulphate levels in river water is typically done through laboratory analysis using methods such as turbidimetry or ion chromatography. The BIS has set a standard value of 200 mg/L for sulphate in river water, which serves as a benchmark for water quality assessment.

Nitrate

Nitrate is a critical parameter used in the prediction of river water quality index. It is an essential nutrient for aquatic plant growth and is naturally present in water bodies. However, excessive amounts of nitrate can lead to eutrophication, which can have harmful effects on aquatic life. The acceptable range for nitrate levels in river water is typically between 2-10 mg/L, with a maximum limit of 50 mg/L to prevent eutrophication. The calculation of nitrate levels in river water is typically done through laboratory analysis using methods such as colourimetry or ion chromatography. The BIS has set a standard value of 10 mg/L for nitrate in river water, which serves as a benchmark for water quality assessment.

Fluoride

Fluoride is a crucial parameter used in the prediction of river water quality index. It is a naturally occurring compound found in rocks and soils, and can also be present in water due to human activities such as industrial discharges and agricultural run-off. While fluoride is essential for dental health, excessive amounts can lead to fluorosis, a condition that can cause discolouration and weakening of teeth and bones. The acceptable range for fluoride levels in river water is typically between 0.5-1.5 mg/L, with a maximum limit of 1.5 mg/L to prevent adverse effects on human health. The calculation of fluoride levels in river water is typically done through laboratory analysis using methods such as ion-selective electrodes. The BIS has set a standard value of 1 mg/L for fluoride in river water, which serves as a benchmark for water quality assessment.

Hardness

Hardness is an essential parameter used in the prediction of water quality index. It is a measure of the amount of dissolved minerals such as calcium and magnesium in the water. High levels of hardness can cause scaling in pipes and appliances and can also affect the effectiveness of cleaning agents. The acceptable range for hardness levels in river water is typically between 50-300 mg/L, with a maximum limit of 600 mg/L to prevent adverse effects on human health and aquatic life. The calculation of hardness levels in river water is typically done through laboratory analysis using methods such as titration or atomic absorption spectrometry. The BIS has set a standard value of 300 mg/L for hardness in river water, which serves as a benchmark for water quality assessment.

Dissolved Oxygen

Dissolved oxygen is a crucial parameter used in the prediction of river water quality index. It is a measure of the amount of oxygen available in water to support aquatic life. DO is essential for the survival of fish, insects, and other organisms living in the water. Low levels of DO can lead to fish kills and other adverse effects on aquatic life. The acceptable range for DO levels in river water is typically between 5-10 mg/L, with a minimum limit of 4 mg/L to prevent adverse effects on aquatic life. The calculation of DO levels in river water is typically done through laboratory analysis using methods such as titration. The BIS has set a standard value of 6 mg/L for DO in river water, which serves as a benchmark for water quality assessment.

Biological Oxygen Demand

Biochemical oxygen demand is a critical parameter used in the prediction of river water quality index. It is a measure of the amount of oxygen required by microorganisms to break down organic matter in the water. High levels of BOD indicate high levels of organic matter in the water, which can lead to reduced DO levels and adverse effects on aquatic life. The acceptable range for BOD levels in river water is typically between 1-6 mg/L, with a maximum limit of 30 mg/L to prevent adverse effects on aquatic life. The calculation of BOD levels in river water is typically done through laboratory analysis using methods such as dilution and incubation. The BIS has set a standard value of 3 mg/L for BOD in river water, which serves as a benchmark for water quality assessment.

Chemical Oxygen Demand

Chemical oxygen demand is an essential parameter used in the prediction of river water quality index. It is a measure of the amount of oxygen required to oxidize organic and inorganic compounds in the water. High levels of COD indicate high levels of pollutants in the water, which can lead to reduced DO levels and adverse effects on aquatic life. The acceptable range for COD levels in river water is typically between 10-30 mg/L, with a maximum limit of 250 mg/L to prevent adverse effects on aquatic life. The calculation of COD levels in river water is typically done through laboratory analysis using methods such as digestion and titration. The BIS has set a standard value of 50 mg/L for COD in river water, which serves as a standard for water quality assessment.

Biological Parameters

Biological water quality parameters are critical indicators used to assess the health and overall quality of water bodies such as rivers, lakes, and streams. These parameters provide information on the presence and abundance of living organisms, such as algae, bacteria, and fish, in the water. The presence of specific species or communities of organisms can be indicative of various water quality characteristics, including nutrient levels, temperature, and dissolved oxygen. The biological parameters such as total coliform and faecal coliform are described below.

Total Coliform

Total Coliform is an important biological parameter used in the prediction of river water quality index. They are a type of bacteria found in the intestines of humans and other warm-

blooded animals, and their presence in water can indicate faecal contamination. High levels of TC in river water can pose a significant risk to public health, as it can be indicative of the presence of harmful pathogens that can cause waterborne illnesses. The acceptable range for TC levels in river water is typically less than 5000 colony-forming units (CFU) per 100 mL, with a maximum limit of 10,000 CFU per 100 mL for potable water. The calculation of TC levels in river water is typically done through laboratory analysis using methods such as membrane filtration and incubation. The BIS has set a standard value of less than 5000 CFU per 100 mL for TC in river water, which serves as a benchmark for water quality assessment.

Faecal Coliform

Faecal Coliform is another important biological parameter used in the prediction of river water quality index. They are a type of bacteria found in the intestines of warm-blooded animals, and their presence in water can indicate faecal contamination from human or animal waste. High levels of FC in river water can pose a significant risk to public health, as it can be indicative of the presence of harmful pathogens that can cause waterborne illnesses. The acceptable range for FC levels in river water is typically less than 1000 CFU per 100 mL, with a maximum limit of 2000 CFU per 100 mL for potable water. The BIS has set a standard value of less than 2500 CFU per 100 mL for FC in river water, which serves as a benchmark for water quality assessment.

Seasonal Parameters

River water quality is subject to significant temporal variability due to changes in hydrological conditions, weather patterns, and anthropogenic activities. Seasonal parameters refer to various climatic and hydrological factors that vary over time. Understanding the temporal variability of water quality is critical for identifying trends, detecting anomalies, and understanding the underlying mechanisms that drive changes in water quality. Therefore, collecting and analysing seasonal parameters are essential for accurate prediction of river WQI and sustainable management of river water resources. Seasonal parameters such as dew, humidity, sea level pressure, precipitation, precip over, wind speed, wind direction, cloud cover, and visibility are considered in this research and are described below.

Dew

Dew plays a crucial role in determining the water quality of rivers, lakes, and groundwater systems. It forms overnight or in the early morning as the air cools, carrying various pollutants like

dust, pollen, and particulate matter. These contaminants can significantly impact the water bodies' health and the well-being of both aquatic ecosystems and humans. Thus, it is imperative to consider dew when predicting the WQI to ensure accurate assessments. However, accurately measuring dew poses challenges, necessitating advanced equipment and precise techniques. Despite these difficulties, incorporating dew in WQI predictions remains vital for the development of dependable models and effective water quality management strategies.

Humidity

Humidity plays a vital role in determining the water quality of rivers, lakes, and groundwater systems. It represents the amount of water vapor in the air and affects the air's capacity to hold pollutants like dust and particulate matter. High humidity levels can facilitate the removal of pollutants from the air, causing them to deposit into water bodies, thus impacting water quality. Additionally, elevated humidity can lead to the formation of fog and dew, further introducing nutrients, metals, and contaminants into water bodies, with significant consequences for aquatic ecosystems and human health. Therefore, including humidity in the prediction of the WQI is crucial for accurate assessments. Fortunately, measuring humidity is relatively straightforward using standard equipment, and its incorporation in WQI predictions aids in developing reliable models and techniques for effective water quality assessment and management.

Sea Level Pressure

Sea level pressure (SLP) is a crucial meteorological parameter that represents the atmospheric pressure at sea level. It serves as a fundamental indicator of weather patterns and atmospheric conditions. SLP influences the movement of air masses and plays a significant role in determining weather systems, including the formation of high and low-pressure areas. These pressure variations impact wind patterns, storm development, and atmospheric circulation, influencing local and regional weather phenomena. Monitoring and analysing SLP data is essential for accurate weather forecasting, climate studies, and understanding the dynamics of atmospheric processes. SLP measurements provide valuable insights into oceanic conditions and their interplay with the atmosphere, aiding in the assessment of marine environments and coastal regions.

Precipitation

Precipitation is a vital component of the Earth's water cycle, representing the various forms of water that fall from the atmosphere to the ground. It includes rain, snow, sleet, and hail, each with its unique characteristics. Precipitation plays a critical role in replenishing water sources, such as rivers, lakes, and groundwater, sustaining ecosystems, and supporting human activities. It is influenced by atmospheric conditions, including temperature, humidity, and air pressure, as well as geographical factors. Precipitation patterns vary across regions and seasons, affecting agriculture, hydroelectric power generation, and overall climate. Monitoring and understanding precipitation trends are crucial for water resource management, weather forecasting, and studying climate change impacts.

Precip Over

Precipitation over a water body significantly impacts its water quality. It can introduce pollutants, nutrients, and sediment into the water, leading to adverse effects on the aquatic ecosystem and human health. Therefore, measuring precipitation over a specific time period is crucial when predicting the WQI of a water body. Collecting precipitation data using standard equipment like rain gauges is feasible, enabling the development of reliable techniques for assessing and managing water quality. Regular and systematic collection of precipitation data is essential due to variations based on location, climate, and season. Including precipitation data in WQI prediction helps identify trends and patterns, facilitating the implementation of effective strategies for water quality management and improvement.

Wind Speed

Wind direction is a critical factor in predicting the WQI of a water body. It has a direct impact on water quality by influencing the movement and distribution of pollutants, nutrients, and sediments within the water column. The direction of wind determines the flow of surface currents, affecting the location and intensity of harmful events like algal blooms. Accurate measurement of wind direction using tools like a wind vane is essential for precise WQI predictions. Regular and systematic collection of wind direction data is crucial. Wind direction data also aids in identifying pollution sources and implementing targeted mitigation measures. Integrating wind direction data into WQI predictions provides valuable insights into the factors impacting water quality and

facilitates the development of practical management practices for ensuring clean and sustainable water resources for both humans and the environment.

Wind Direction

Wind direction is an important factor to consider when predicting the WQI of a water body. Wind direction influences water quality by affecting the transport and distribution of pollutants, nutrients, and sediments within the water column. Measuring wind direction is crucial to predicting the WQI accurately. Wind direction can be measured using instruments such as a wind vane, and data collected regularly and systematically. By incorporating wind direction data into models and techniques for WQI prediction, water quality assessments can be improved, leading to effective management strategies for maintaining and improving water quality. Wind direction data can also help identify potential sources of pollution and inform targeted management practices to address these issues. Incorporating wind direction data into WQI prediction can provide valuable insights into the factors that influence water quality and help develop effective management practices to ensure clean and healthy water for human and environmental well-being.

Cloud Cover

Cloud cover is a significant parameter that influences the WQI of a water body. It impacts water quality by affecting light availability for photosynthesis, thereby influencing the growth and distribution of aquatic plants and algae. Moreover, cloud cover also affects the temperature and heat balance of the water body, influencing chemical and biological processes that impact water quality. Accurate measurement of cloud cover is essential for precise WQI predictions. Cloud cover data can be obtained through satellite or ground-based observations, collected regularly and systematically. Incorporating this data into WQI models enhances water quality assessments and enables effective management strategies to maintain and improve water quality. Cloud cover data can reveal seasonal or regional patterns, aiding in targeted management practices. Integrating cloud cover data into WQI predictions provides valuable insights into the factors influencing water quality and supports the development of practical management practices to ensure clean and sustainable water resources for both humans and the environment.

Visibility

Visibility plays a crucial role in predicting the WQI of a water body. It refers to the distance at which objects can be clearly seen in the water and is influenced by factors such as suspended

particles, dissolved substances, and algae. A decrease in visibility indicates higher levels of pollutants, which can have detrimental effects on aquatic life and human health. Therefore, measuring visibility is essential for accurate WQI predictions. Remote sensing and satellite imagery have proven effective in assessing visibility and its impact on water quality. Implementing management practices to improve visibility, such as reducing sediment runoff and controlling nutrient inputs, can contribute to a healthier aquatic ecosystem. By considering visibility in WQI predictions, can gain valuable insights into the overall water quality and implement targeted measures to safeguard water resources for the well-being of both the environment and society.

The physical, chemical and biological observations are collected from sampling stations of both rivers and the seasonal parameters are collected from visual crossing site based on locations of each monitoring stations. The total number of instances obtained for Bhavani and Bharathapuzha rivers are 10560 and 2190 respectively. The list of parameters identified and collected for the study is illustrated in Table III. The sample physicochemical parameter data collected from Bhavani River is given in Table IV, and the pooled feature Bhavani River features are shown in Table V. The sample data of Bharathapuzha River is illustrated in Table VI.

Table III. List of Water Quality Parameters

Physicochemical Parameters		**Seasonal Parameters**
pH	TSS	Temperature
Conductivity	TDS	Dew
Turbidity	FDS	Humidity
Phenolphthalein Alkalinity	Phosphate	Sea level pressure
Total Alkalinity	Boron	Precipitation
Chloride	Potassium	Precip cover
COD	BOD	Windspeed
TKN	Fluoride	Wind dir
Ammonia	Nitrate-N	Cloud cover
Hardness	TC	Visibility
Ca. hardness	FC	**Spatial Parameters**
Mg. hardness	Dissolved Oxygen	Station ID
Sulphate	**Temporal Parameter**	Latitude
Sodium	Date	Longitude

Table IV. Sample Bhavani River Data - Physicochemical Parameters

pH	7.15	7.46	7.5	7.18	7.45	7.05	7.4	7.38	7.56	7.1
Conductivity	340	339	339	340	340	342	341	339	340	340
Turbidity	2	2	2	2	2	2	2	2	2	2
Total Alkalinity	111	110	112	111	110	110	112	111	112	111
Chloride	21	21	22	21	20	20	20	21	21	21
COD	4	3.9	4	3.9	4	4	4	3.9	3.9	4
TKN	0.1	0.1	0.09	0.1	0.1	0.09	0.1	0.1	0.1	0.11
Ammonia	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Hardness	118	118	119	119	119	119	118	118	118	117.5
Ca. hardness	74	74	74.5	74.5	74	73.5	73.5	73.5	74	74
Mg. Hardness	44	44	44	43.5	43.5	43.5	44	44	44	44
Sulphate	12	12.5	12	12	12.5	12.5	12	12	12.5	12
Sodium	27.1	27.1	27.2	27.2	27	27.1	27.1	27	27.1	27.1
TSS	300	300	300	300	300	300	300	300	300	300
TDS	190	190	189	189	189	190	189	190	189	188
FDS	174	174	174	174.5	174.5	174	174	174	173.5	173
Phosphate	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
Potassium	2.67	2.67	2.66	2.66	2.67	2.67	2.66	2.66	2.66	2.66
BOD	0.89	0.87	0.89	0.88	0.85	0.87	0.82	0.81	0.88	0.82
Fluoride	0.12	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.18	0.18
Nitrate-N	1.1	1.1	1.1	1	1.2	1	1.2	1.2	1.2	1.2
DO	6.99	6.97	6.81	7.19	7.3	7.39	7.06	7.02	6.97	7.39
TC	88	98	118	86	65	105	83	113	65	85
FC	80	80	80	79.5	79.5	79	79.5	80	80	80

Table V. Sample Bhavani River Data- Physiochemical and Seasonal Parameters

pH	7.15	7.46	7.5	7.18	7.45	7.05	7.4	7.38	7.56	7.1
Conductivity	340	339	339	340	340	342	341	339	340	340
Turbidity	2	2	2	2	2	2	2	2	2	2
Total Alkalinity	111	110	112	111	110	110	112	111	112	111
Chloride	21	21	22	21	20	20	20	21	21	21
COD	4	3.9	4	3.9	4	4	4	3.9	3.9	4
TKN	0.1	0.1	0.09	0.1	0.1	0.09	0.1	0.1	0.1	0.11
Ammonia	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Hardness	118	118	119	119	119	119	118	118	118	117.5
Ca. hardness	74	74	74.5	74.5	74	73.5	73.5	73.5	74	74
Mg. Hardness	44	44	44	43.5	43.5	43.5	44	44	44	44
Sulphate	12	12.5	12	12	12.5	12.5	12	12	12.5	12
Sodium	27.1	27.1	27.2	27.2	27	27.1	27.1	27	27.1	27.1
TSS	300	300	300	300	300	300	300	300	300	300
TDS	190	190	189	189	189	190	189	190	189	188
FDS	174	174	174	174.5	174.5	174	174	174	173.5	173
Phosphate	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
Potassium	2.67	2.67	2.66	2.66	2.67	2.67	2.66	2.66	2.66	2.66
BOD	0.89	0.87	0.89	0.88	0.85	0.87	0.82	0.81	0.88	0.82
Fluoride	0.12	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.18	0.18
Nitrate-N	1.1	1.1	1.1	1	1.2	1	1.2	1.2	1.2	1.2
DO	6.99	6.97	6.81	7.19	7.3	7.39	7.06	7.02	6.97	7.39
TC	88	98	118	86	65	105	83	113	65	85
FC	80	80	80	79.5	79.5	79	79.5	80	80	80
Temp	25	24	25	25	25	24	24	25	25	25
Dew	15.7	14.6	13.4	13.6	15.6	17.7	18.9	19.4	18.3	17.8
Humidity	59.3	56.72	51.89	53.06	58.8	62.79	68.91	68.63	65.71	63.8
Sea level pressure	1016.6	1017.1	1015.8	1015.7	1014.8	1014.8	1015.5	1015.5	1013.7	1014.5
Precipitation	0	0	0	0	0	0	0.2	0	0	0
Precip cover	0	0	0	0	0	0	4.17	0	0	0
Wind speed	16.3	14.4	13.1	15.4	14	18.7	40.2	13.6	14.4	14.9
Wind dir	52.9	62.3	61.7	68.2	56.5	69.3	114.6	95	94.9	65.1
Cloud cover	27.4	17.9	5.5	14.1	14.6	16	32.3	42.5	26.3	14
Visibility	5.5	6	5.7	5.9	5.6	5.5	4.8	5.3	5.1	5.4

Table VI. Sample Bharathapuzha River Data-Physiochemical and Seasonal Parameters

Temp	29	27	30	28	30	30	30	28	29	29
pH	7.31	7.60	7.43	7.27	7.63	7.20	7.52	7.47	7.38	7.34
Conductivity	312	316	295	290	316	293	295	308	312	293
Turbidity	2	2	2	2	2	2	2	2	2	2
Total Alkalinity	77	75	76	79	80	78	79	76	79	80
Chloride	45	43	39	38	39	39	42	41	43	43
COD	4	4	4	4	4	4	4	4	4	4
TKN	0.1	0.1	0.09	0.1	0.1	0.09	0.1	0.1	0.1	0.11
Ammonia	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Hardness	105	100	104	108	103	107	105	105	108	107
Ca. hardness	54	50	51	53	50	50	51	53	51	50
Mg. Hardness	53	51	50	50	52	50	52	52	52	53
Sulphate	0.10	0.31	0.14	0.28	0.31	0.18	0.07	0.21	0.58	0.60
Sodium	16	15	17	17	18	15	15	15	17	18
TDS	180	173	171	175	180	177	180	172	182	179
FDS	99	101	94	102	101	100	100	95	94	102
TSS	300	300	300	300	300	300	300	300	300	300
Phosphate	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Potassium	6.38	6.40	6.46	6.42	6.45	6.41	4.10	4.25	4.45	4.21
BOD	1.01	1.03	1.02	1.02	1.03	1.02	1.00	1.03	1.00	1.00
Fluoride	0.58	0.45	0.50	0.45	0.46	0.56	0.43	0.37	0.41	0.33
Nitrate-N	1.29	1.03	1.51	1.19	1.78	1.49	0.92	1.29	1.10	1.76
TC	320	306	338	312	261	338	323	293	313	311
FC	281	272	258	264	254	296	265	254	265	260
DO	6.12	6.25	6.12	6.25	6.24	6.18	6.26	6.06	6.06	6.07
Dew	15.7	14.6	13.4	13.6	15.6	17.7	18.9	19.4	18.3	17.8
Humidity	59.30	56.72	51.89	53.06	58.80	62.79	68.91	68.63	65.71	63.80
SLP	1016.6	1017.1	1015.80	1015.70	1014.80	1014.80	1015.50	1015.50	1013.70	1014.50
Precipitation	0	2	0	4	18	19.41	14.81	14	5	11.85
Precipcover	0	4.17	0	8.33	4.17	8.33	8.33	4.17	8.33	8.33
Windspeed	16.3	14.4	13.1	15.4	14	18.7	40.2	13.6	14.4	14.9
Winddir	52.9	62.3	61.7	68.2	56.5	69.3	114.6	95	94.9	65.1
Cloud cover	27.4	17.9	5.5	14.1	14.6	16	32.3	42.5	26.3	14
Visibility	5.5	6	5.7	5.9	5.6	5.5	4.8	5.3	5.1	5.4

3.3 COMPUTATION OF WQI

The WQI represents the overall impact of water quality standards on water quality. Its objective is to translate complex water quality data into meaningful and easily understood information. In order to evaluate water quality, parameters must be selected in accordance with a set standard, such as the Indian Standard for Drinking Water Specification (BIS 2004). The BIS is the national standards organization of India and has established a standard for the calculation of the WQI. The BIS standard for the WQI uses a multi-attribute index approach, which considers multiple water quality parameters and assigns weights to each parameter based on its relative importance. The resulting index provides an overall assessment of water quality and is used to evaluate the suitability of water for different uses such as drinking, irrigation, and recreation.

The WQI provides a comprehensive evaluation of water quality by taking into account various physical, chemical, and biological parameters. The WQI is calculated using a weighted average of water quality parameters like pH, dissolved oxygen, turbidity, faecal coliform, and etc. The computation of the WQI involves the following steps:

- Assigning weights to each water quality parameter based on their relative importance and the weighted are calculated using the formula: $W_n = K/S_n$, where W_n is the relative weight, K is the weight of each parameter, and S_n is the permissible limit.
- Assigning a water quality rating (Q_n) for each parameter:

  $Q_n = (V_n / S_n) \times 100$, where Q_n is the water quality rating, Vn is the mean concentration value for each parameter, and S_n is the desirable limit as specified in the BIS 2004 Indian drinking water standard.
- Determining the sub-index (SI) for each water quality parameter:

  $W_n \times Q_n$ SI, where SI is the sub-index of the parameter and Q_n is the rating based on parameter concentration.
- Calculating the WQI by summing the SI of each water quality parameter.

  $$WQI = \frac{\Sigma WnQn}{\Sigma Wn}$$

The characteristics of water quality parameters are analysed in accordance with the BIS drinking water quality requirements, as outlined in Table VII, which displays the BIS water quality parameters permissible limits and the formula used to calculate the WQI.

Table VII. Water Quality Parameter for Computing WQI

Parameters	S_n	$1/S_n$	$\sum 1/S_n$	K	W_n	Ideal Value	V_n	V_n/S_n	Q_n	W_n*Q_n
Temp	28	0.04	8.4048	0.12	0	0	28	0.4	40	0.17
pH	8.5	0.12	8.4048	0.12	0.01	7	7.3	0.86	85.88	1.2
Conductivity	150	0.01	8.4048	0.12	0	0	65	0.43	43.33	0.03
Turbidity	5	0.2	8.4048	0.12	0.02	0	2	0.4	40	0.95
PA	20	0.05	8.4048	0.12	0.01	0	0	0	0	0
TA	200	0.01	8.4048	0.12	0	0	10	0.05	5	0
Chloride	250	0	8.4048	0.12	0	0	10	0.04	4	0
COD	10	0.1	8.4048	0.12	0.01	0	4	0.4	40	0.48
TKN	100	0.01	8.4048	0.12	0	0	0.19	0	0.19	0
Ammonia	50	0.02	8.4048	0.12	0	0	0.25	0.01	0.5	0
Hardness	100	0.01	8.4048	0.12	0	0	9	0.09	9	0.01
Ca. Hardness	75	0.01	8.4048	0.12	0	0	7	0.09	9.33	0.01
Mg. Hardness	30	0.03	8.4048	0.12	0	0	6	0.2	20	0.08
Sulphate	200	0.01	8.4048	0.12	0	0	3	0.02	1.5	0
Sodium	200	0.01	8.4048	0.12	0	0	7	0.04	3.5	0
TSS	300	0	8.4048	0.12	0	0	300	1	100	0.04
TDS	1000	0	8.4048	0.12	0	0	45	0.05	4.5	0
FDS	200	0.01	8.4048	0.12	0	0	35	0.18	17.5	0.01
Phosphate	0.3	3.33	8.4048	0.12	0.4	0	0.11	0.37	36.67	14.54
Boron	1	1	8.4048	0.12	0.12	0	0.1	0.1	10	1.19
Potassium	2.5	0.4	8.4048	0.12	0.05	0	2	0.8	80	3.81
BOD	3	0.33	8.4048	0.12	0.04	0	2.3	0.77	76.67	3.04
Fluoride	1.5	0.67	8.4048	0.12	0.08	0	0.119	0.08	7.93	0.63
DO	7.5	0.13	8.4048	0.12	0.02	14	8.3	1.22	121.82	1.93
Nitrate-N	0.503	1.99	8.4048	0.12	0.24	0	0.902	1.79	179.32	42.42
TC	100	0.01	8.4048	0.12	0	0	60	0.6	60	0.07
FC	60	0.02	8.4048	0.12	0	0	44	0.73	73.33	0.15
WQI										70.76

In selecting the water quality parameters and determining the range for the Water Quality Index, it is crucial to consider the impact of each parameter on water quality and the potential health implications. The range of the WQI serves as a guide to interpreting water quality data, highlighting the significance of each WQI value and the corresponding water quality conditions.

The Water Quality Index is a numerical representation of water quality that ranges from 0 to 121. The higher the index value, the lower the quality of water. A WQI of 121 or above is considered unacceptable and is categorized as Class E. Water with an index between 91 and 120 is considered very poor and is classified as Class D. An index between 61 and 90 is considered poor and is classified as Class C. A WQI between 31 and 60 is considered good and is categorized as Class B. On the other hand, a WQI of 0 to 30 is considered excellent and is classified as Class A. This index provides a clear and easily understandable way of evaluating the suitability of water for various purposes, such as drinking, irrigation which is depicted in Table VIII.

Table VIII. BIS (2004) Water Quality Standards

Water Quality Index Value	**Water Quality Index Class**	**Water Quality Label**
>121	E	Unsuitable
91-120	D	Very Poor
61-90	C	Poor
31-60	B	Good
0-30	A	Excellent

For each sample of the time series river data, the WQI values are calculated and assigned to the respective samples as the target variable to facilitate the regression modelling. Once the data collection is completed, the EDA is done to identify patterns trends and potential outliers in the data. The distribution of observations and the statistical characteristics of the data collected from monitoring stations regarding the water quality parameters are analysed using explorative and descriptive statistical analysis.

3.4. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a crucial step in the data mining process that involves examining and understanding the characteristics of a dataset. It serves as an initial exploration to gain insights, discover patterns, and detect anomalies in the data before applying more advanced data mining techniques. EDA aims to uncover relationships, identify trends, and extract meaningful information from the data, providing a solid foundation for further analysis and decision-making. EDA plays a pivotal role in understanding the structure of data, assessing its quality, and generating valuable insights that can drive meaningful conclusions and actions.

EDA TECHNIQUES

Several techniques are available to carry out EDA for a better understanding of the data before applying any modelling techniques. Some common EDA techniques are:

Univariate Analysis: This technique is used to analyse a single variable in isolation. It involves visualizing and summarizing the data using histograms, box plots, and summary statistics like mean, median, and mode.

Bivariate Analysis: This technique is used to analyse the relationship between two variables. It involves visualizing and summarizing the data using scatterplots, correlation analysis, and regression analysis.

Multivariate Analysis: This technique is used to analyse the relationship between three or more variables. It involves visualizing and summarizing the data using scatterplots, correlation analysis, and regression analysis.

Data Visualization: This technique is used to present data in a graphical format, making it easier to understand and interpret. Common visualization techniques include bar charts, line charts, scatterplots, heat maps, and box plots.

Outlier Detection: This technique is used to identify and handle outliers in the data. Outliers are data points that are significantly different from the rest of the data and can have a significant impact on the analysis results.

Missing Value Analysis: This technique is used to identify missing values in the data and to handle them appropriately. Missing values can have a significant impact on the analysis results and can cause biases if not handled correctly.

Dimensionality Reduction: This technique is used to reduce the number of variables in the data by identifying and removing redundant or irrelevant variables. Common techniques include PCA and factor analysis.

Statistical charts such as box plot, pair plot, heat map, and bar graph are commonly used to visually explore and summarize the data. Box plot is used to display the distribution of a variable, including outliers and quartiles. Pair plot is used to visualize the pairwise relationships between variables in a dataset, making it easy to identify patterns and correlations. Heat map is used to show the relationship between two variables using a colour-coded matrix. The bar graph is used to display the comparison between different categories or groups. Explorative and descriptive analysis of the Bhavani River data and Bharathapuzha data are described below.

Box Plot Analysis

Box plot analysis, also known as box-and-whisker plot analysis, is a graphical representation of the distribution of a dataset, providing a visual summary of key statistical measures such as median, quartiles, and potential outliers. It is found that the temperature values lie between 22 and 33, but most values lie between 25 and 28. Biological oxygen demand values range from 0 to 2, with the majority of values falling between 0.8 and 1.8. In most cases, values range from 65 to 150, and total alkalinity values range from 804 to 1, with most cases falling between 45 and 98. Conductivity has a wide range of values, i.e., from 1 to 1200, but most values lie between 60 and 210. Similarly, for total coliform, the values lie from 10 to 2500, with most values lying in the range of 10 to 300. It is observed that the parameters conductivity, total coliform has a wide range value than other parameters. Hence the parameters need to be normalized during pre-processing for building an efficient time-series dataset. Fig.3.4 depicts a box plot analysis of physicochemical parameters from Bhavani River for some meaningful variables related to river water quality, such as conductivity, temperature, pH, turbidity, nitrate, TC, TDS, sulphate, BOD, COD, dissolved oxygen. Box plot analysis of the Bhavani River with seasonal variables are illustrated in Fig. 3.5. The box plot analysis of Bharathapuzha River with both physiochemical and seasonal parameters are shown in Fig. 3.6.

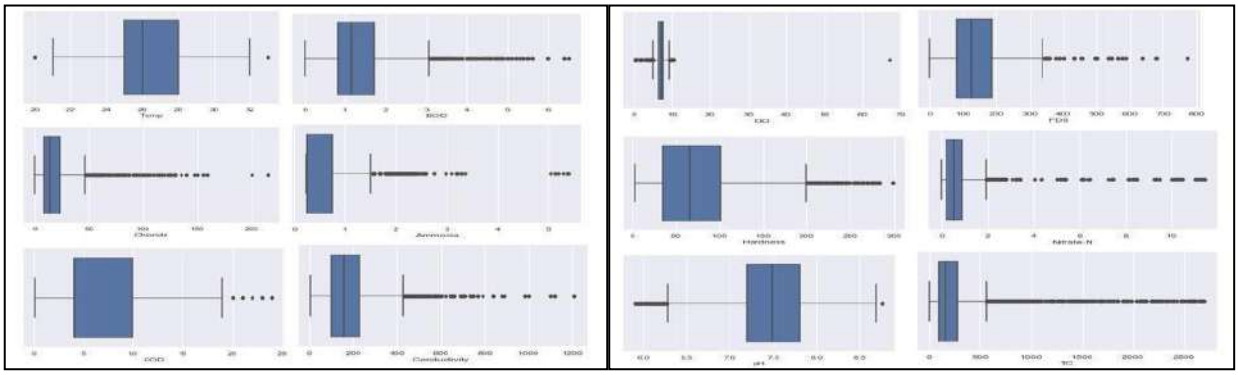

Fig. 3.4 Sample Box Plot Visualization of Physiochemical Parameters of Bhavani River

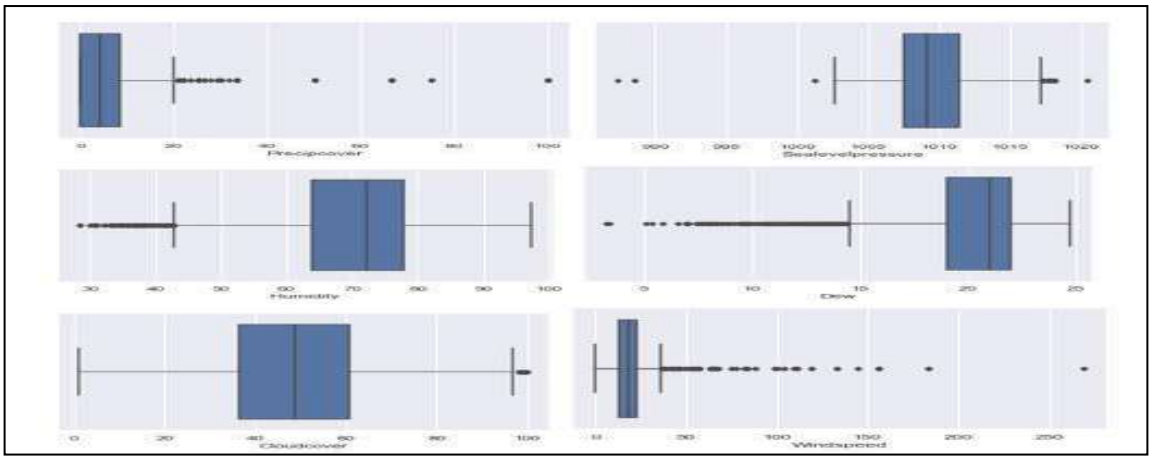

Fig. 3.5. Sample Box Plot Visualization of Seasonal Parameters of Bhavani River

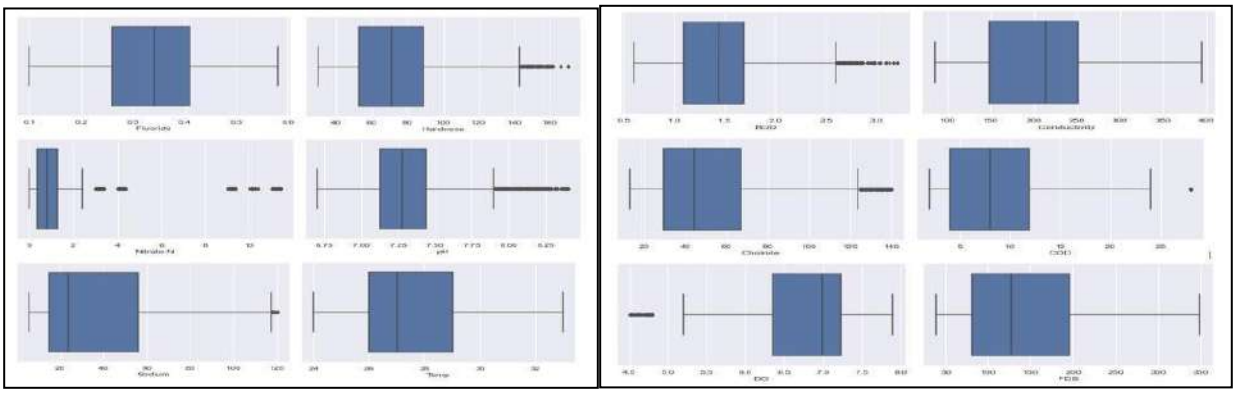

Fig. 3.6. Sample Box Plot Visualization of Pooled Parameters of Bharathapuzha River

Histogram Analysis

Histogram analysis of ammonia, BOD, COD, conductivity, DO, FDS, fluoride, hardness, nitrate, potassium, sodium, TC, TDS, TKN, temperature, total alkalinity, turbidity, and pH is done using the same river quality dataset. The histogram representation helps in understanding the range of values, like the pH value being between six and eight. Some parameters, like chloride, conductivity, FDS, hardness, sodium, TC, TKN, and total alkalinity, have a wide range of values. The minimum value of chloride is zero and the maximum value is 215, whereas the conductivity maximum value is 400 and the minimum value is 6.4. The total coliform minimum value is 10 and the maximum value is 2500, which shows that the range of parameter values is large. Hence the parameters need to be standardized in values to fall within a range. Fig.3.7 depicts the histogram analysis of different physiochemical parameters of Bhavani River whereas Fig.3.8 illustrates the analysis of pooled parameters of Bhavani River and the analysis of pooled parameters of Bharathapuzha River is shown in Fig.3.9.

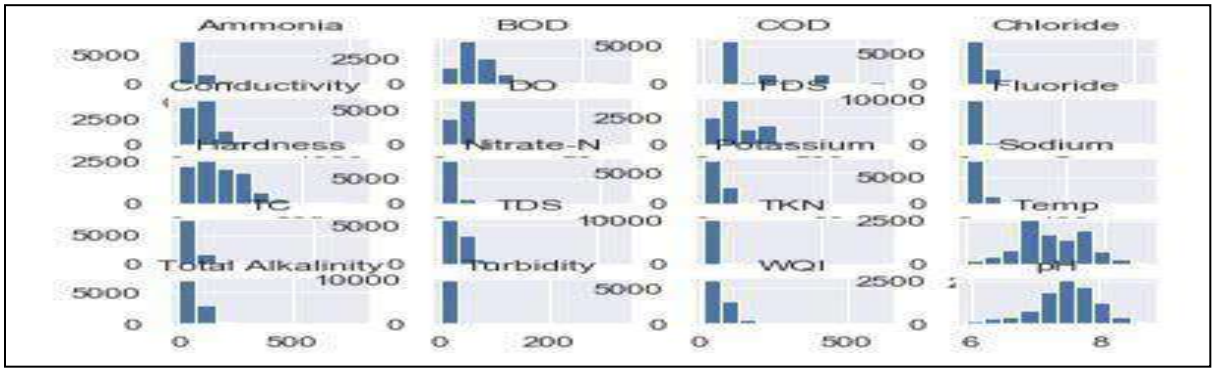

Fig.3.7. Sample Histogram Visualization of Physicochemical Parameters of Bhavani River

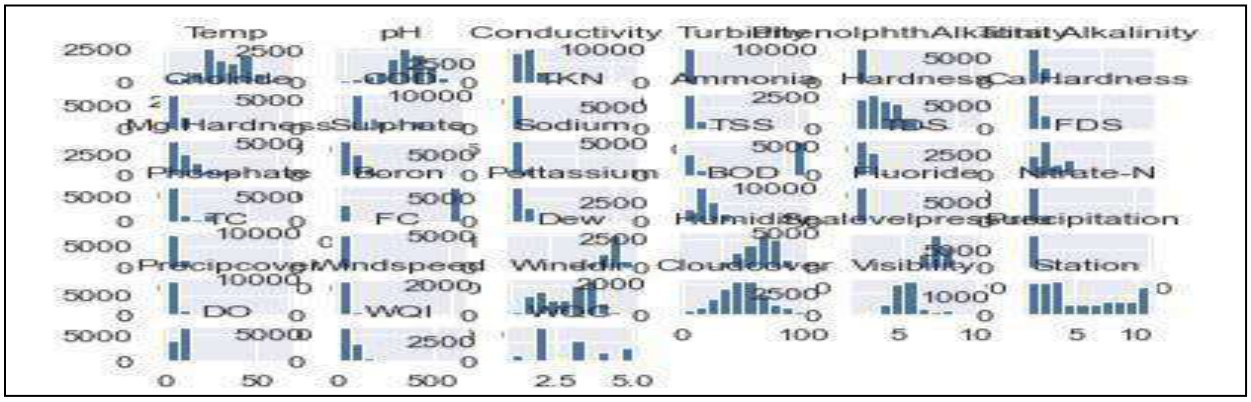

Fig. 3.8. Sample Histogram Visualization of Pooled Parameters of Bhavani River

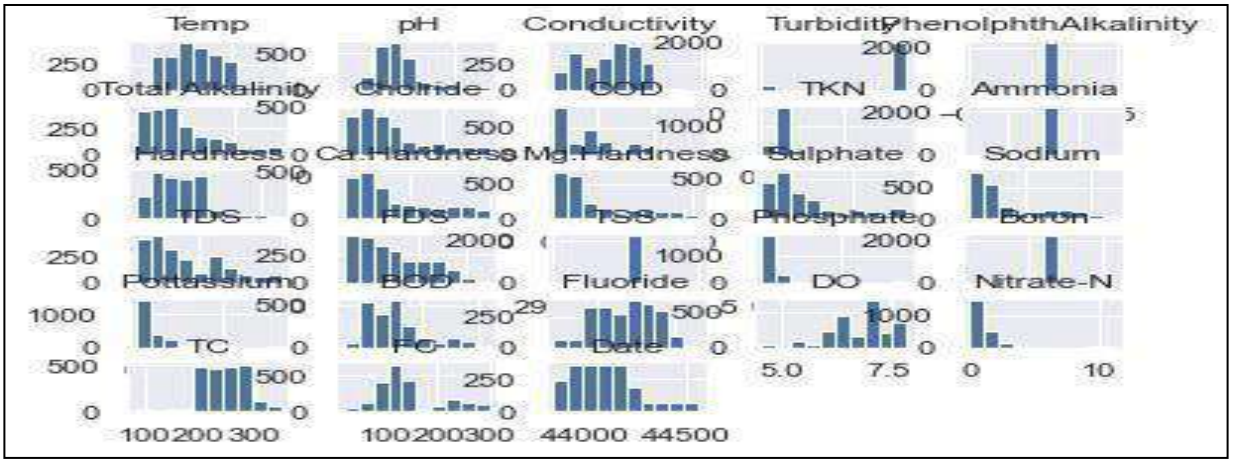

Fig. 3.9. Sample Histogram Visualization of Pooled Parameters of Bharathapuzha River

Pair Plots Analysis

A pair plot depicts all possible relations between each parameter in the river water quality dataset. Relationships between each parameter are visualized using bar graphs and scatter plots. The figure depicted below shows how temperature is correlated with pH, conductivity, total alkalinity, and other attributes. If the plot is scattered, then the correlation is less. For instance, the water quality index is highly correlated with nitrate and it is negatively correlated with DO. The bar graph in the pair plot shows the range of values within which it lies and how many instances lie within the range. A bar graph of pH values shows that it ranges from 6 to 8.5, with the majority of instances falling within the range of eight. The scatter plots of temperature show a high correlation with pH, chloride, and chemical oxygen demand and less correlation with nitrate, dissolved oxygen, FDS, and turbidity. A pair plot for some meaningful variables related to river water quality, such as conductivity, temperature, pH, turbidity, nitrate, TC, TDS, sulphate, BOD, COD, dissolved oxygen, and seasonal parameters. The pair plot analysis of physiochemical parameters of Bhavani River is depicted in Fig.3.10, whereas the pair plot analysis of the pooled parameters of Bhavani River is depicted in Fig.3.11 and the pair plot analysis of the pooled parameters of Bharathapuzha River is shown in Fig. 3.12.

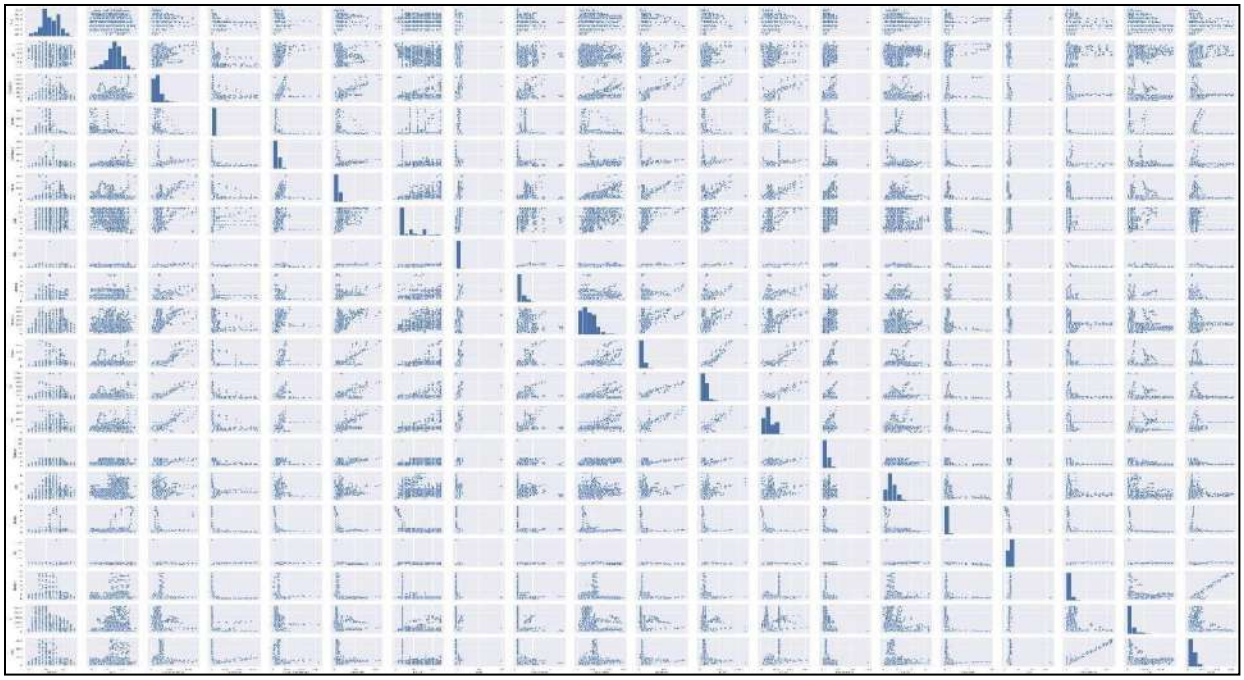

Fig.3.10. Sample Pair plot Visualization of Physiochemical Parameters of Bhavani River

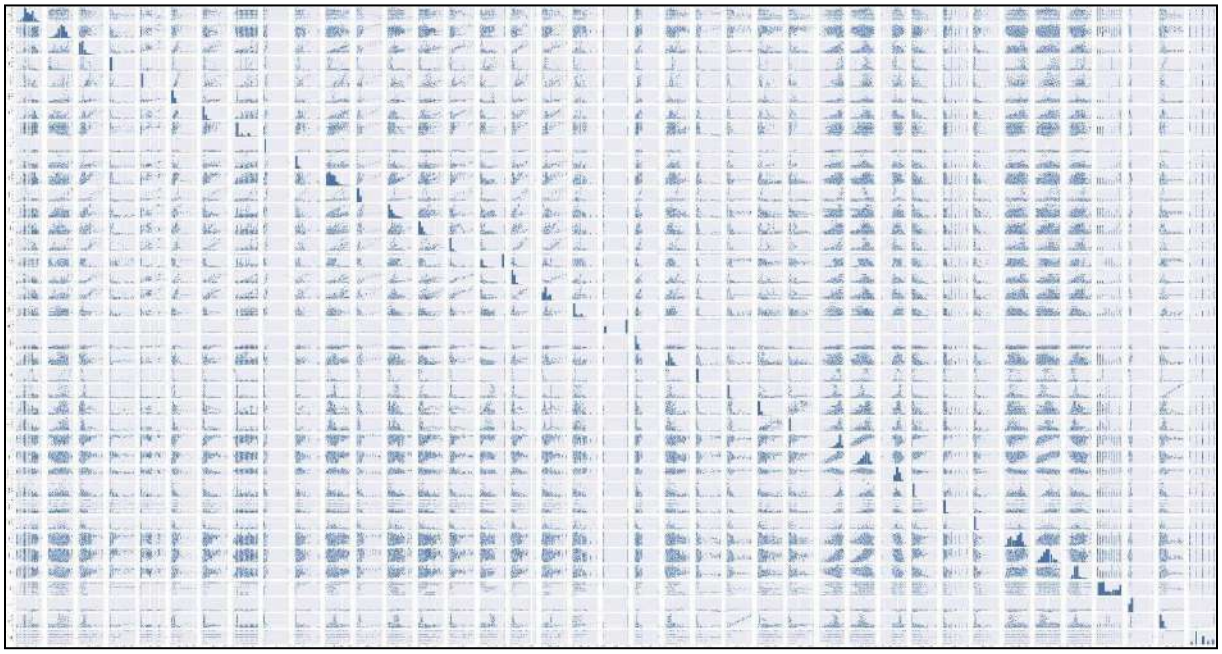

Fig. 3.11. Sample Pair plot Visualization of Pooled Parameters of Bhavani River

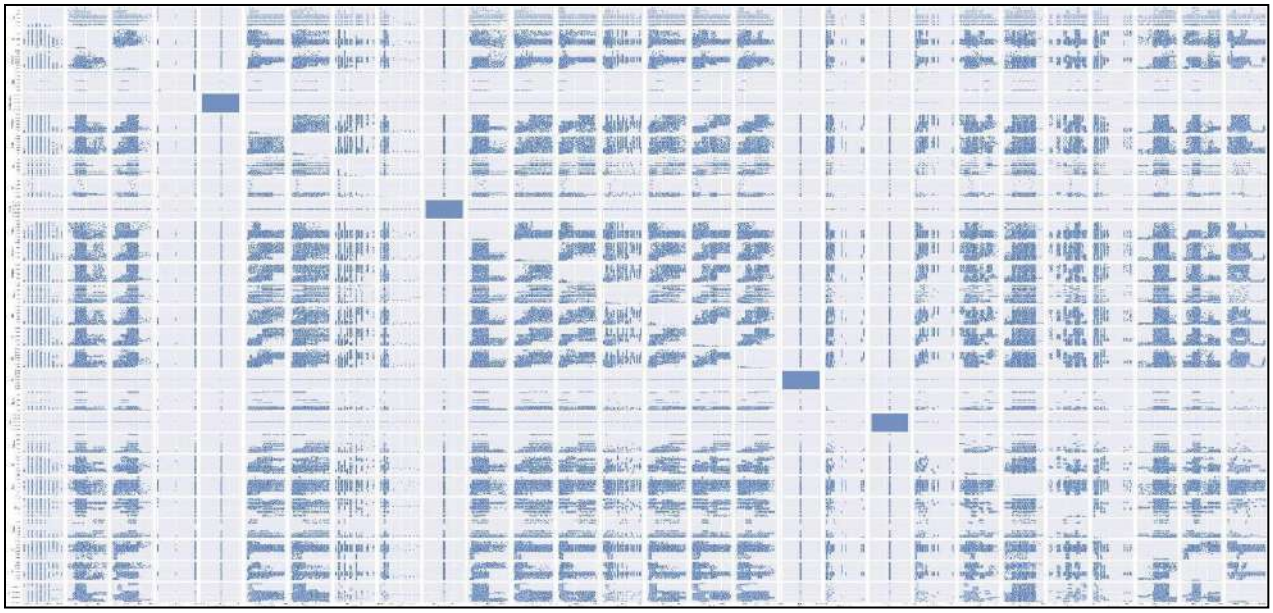

Fig. 3.12. Sample Pair plot Visualization of Pooled Parameters of Bharathapuzha River

Heat Map Analysis

A heat map is the visual representation of the correlation matrix. Temperature is negatively correlated with pH, turbidity, FDS, TC, DO, and nitrate, whereas other attributes are positively correlated. pH is negatively correlated with temp, turbidity, chloride, COD, TKN, ammonia, sodium, BOD, and potassium. For other parameters, it is positively correlated. Conductivity, hardness, and alkalinity are negatively correlated only with DO, with all other attributes being positively correlated. pH, DO, and nitrate are negatively correlated with pH, DO, and nitrate, whereas all other attributes are positively correlated. Turbidity is negatively correlated with temperature, pH, DO, and nitrate. TDS is positively related to all attributes except DO and nitrate. FDS is negatively correlated with temp, turbidity, fluoride, and DO. All other attributes are positively correlated. BOD is negatively correlated with pH, DO, nitrate, and TC, whereas fluoride is negatively correlated with FDS, DO nitrate, and TC. DO is positively correlated with only pH and nitrate. TC is negatively correlated with temperature, potassium, fluoride, and DO. Nitrate is positively correlated with pH, conductivity, total alkalinity, hardness, FDS, DO, and TC. The heat map analysis of the physiochemical parameters of Bhavani River is illustrated in Fig.3.13 and the pooled parameters of Bhavani River is shown in Fig.3.14. The heat map analysis of the Bharathapuzha river for the pooled parameters is depicted in Fig.3.15.

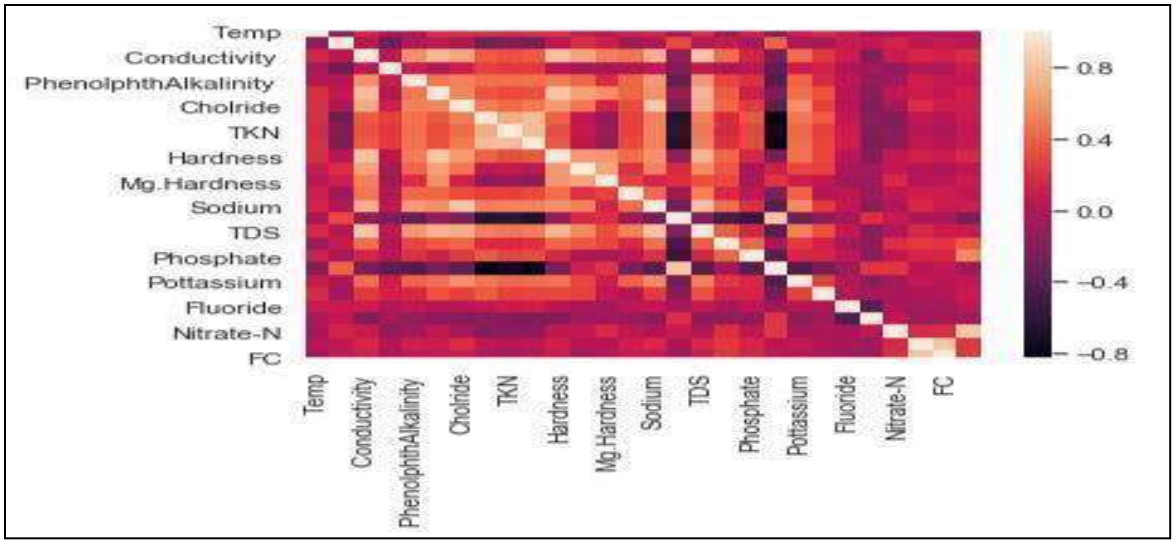

Fig. 3.13. Sample Heat Map Visualization of Physiochemical Parameters of Bhavani River

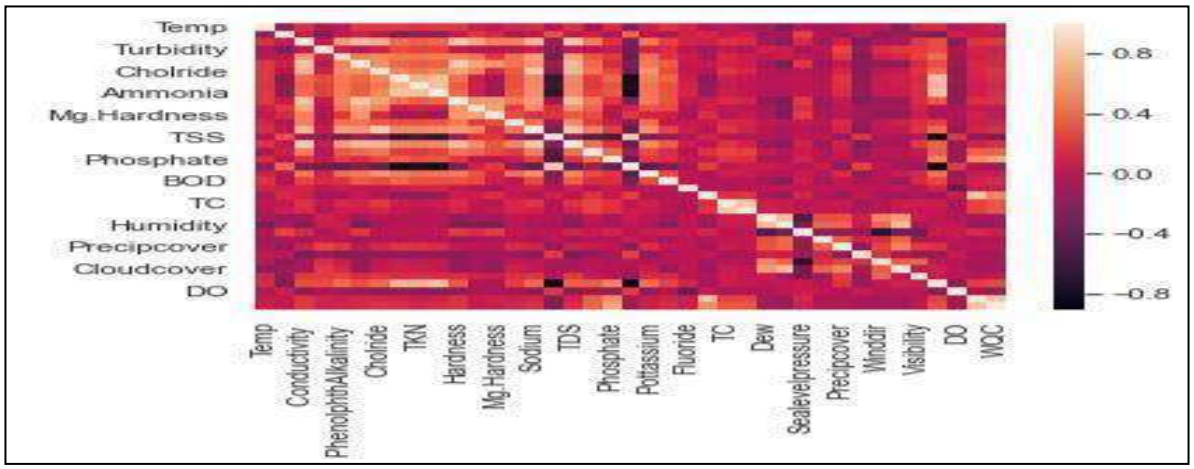

Fig.3.14. Sample Heat Map Visualization of Pooled Parameters of Bhavani River

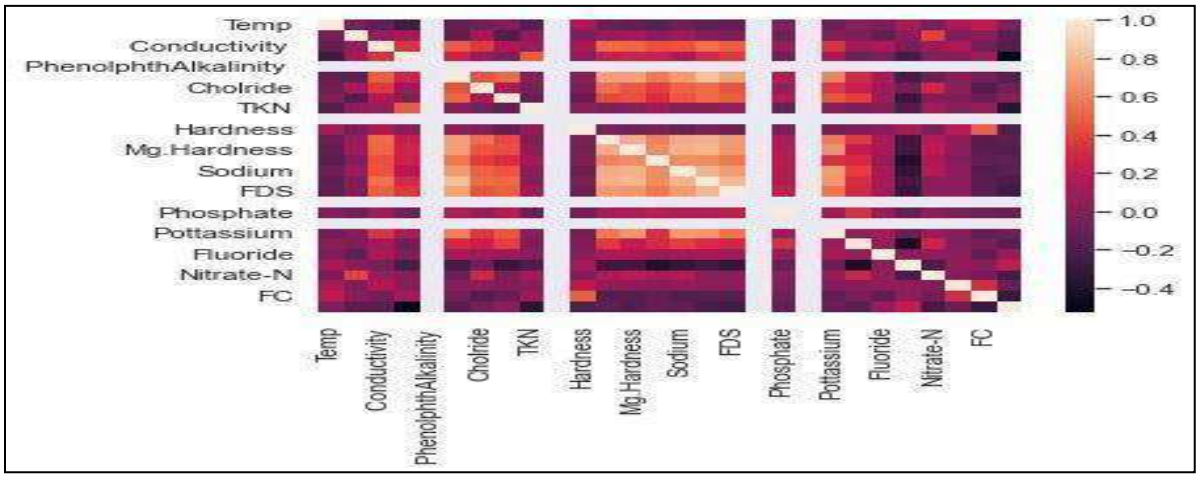

Fig. 3.15. Sample Heat Map Visualization of Pooled Parameters of Bharathapuzha River

TEMPORAL VARIATION OF PARAMETERS

Temporal variation of physiochemical and seasonal parameters is an important aspect to consider while predicting the WQI of a river. The physiochemical parameters such as dissolved oxygen, pH, temperature, conductivity, and turbidity changes over time due to seasonal variations, natural processes, or anthropogenic activities. Similarly, seasonal parameters such as rainfall, temperature, and land use also affect the water quality of a river. The temporal variation of physiochemical and seasonal parameters of both Bhavani River and Bharathapuzha River is carried out. From the analysis it is identified that how each parameter influences the water quality. The significance of temporal variation of physiochemical parameters with WQI of Bhavani River is shown in Fig.3.16 and the variation of seasonal parameters is depicted in Fig.3.17. The temporal variation of pooled parameters of Bharathapuzha River is illustrated in Fig 3.18.

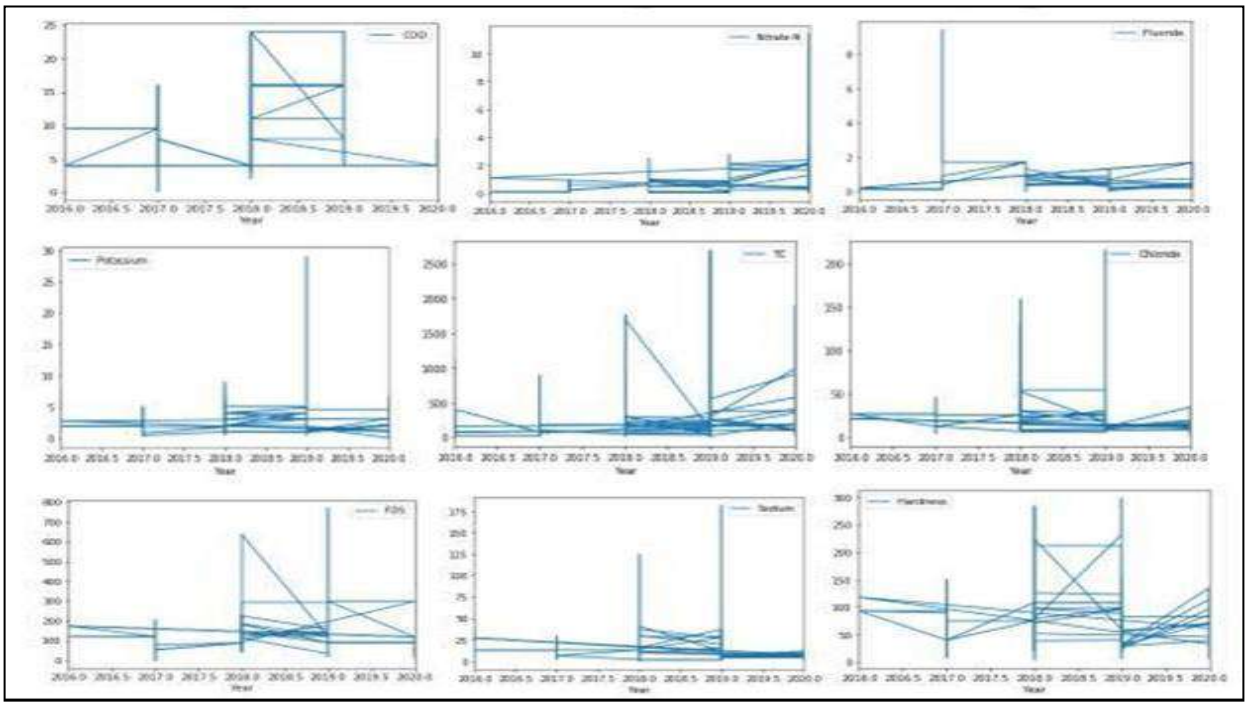

Fig. 3.16. Sample Temporal Variation of Physiochemical Parameters of Bhavani River

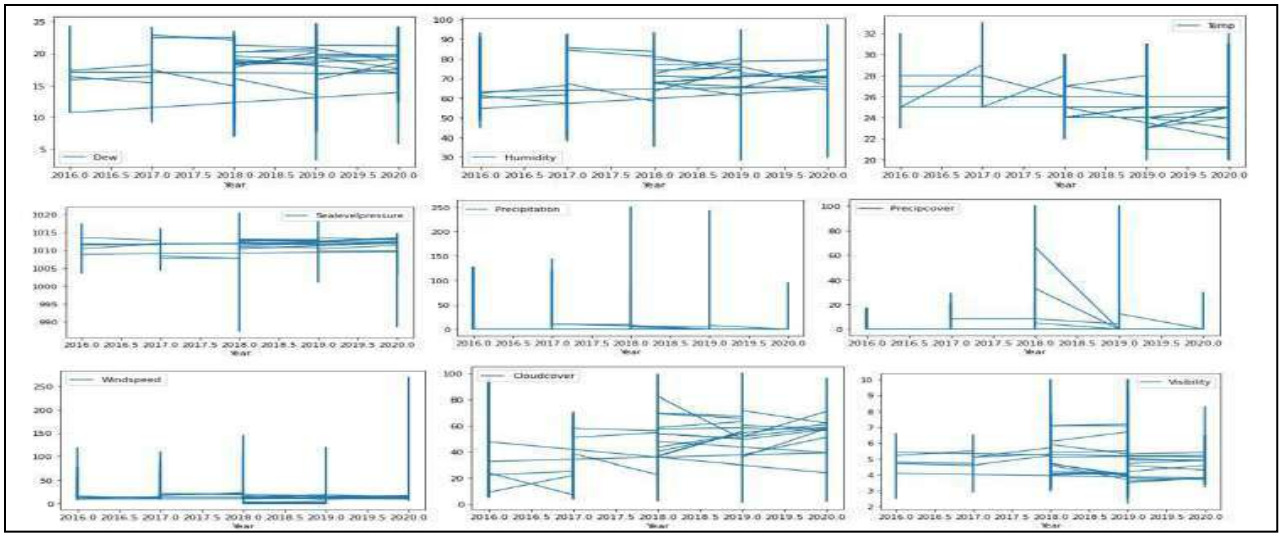

Fig.3.17. Sample Temporal Variation of Seasonal Parameters of Bhavani River

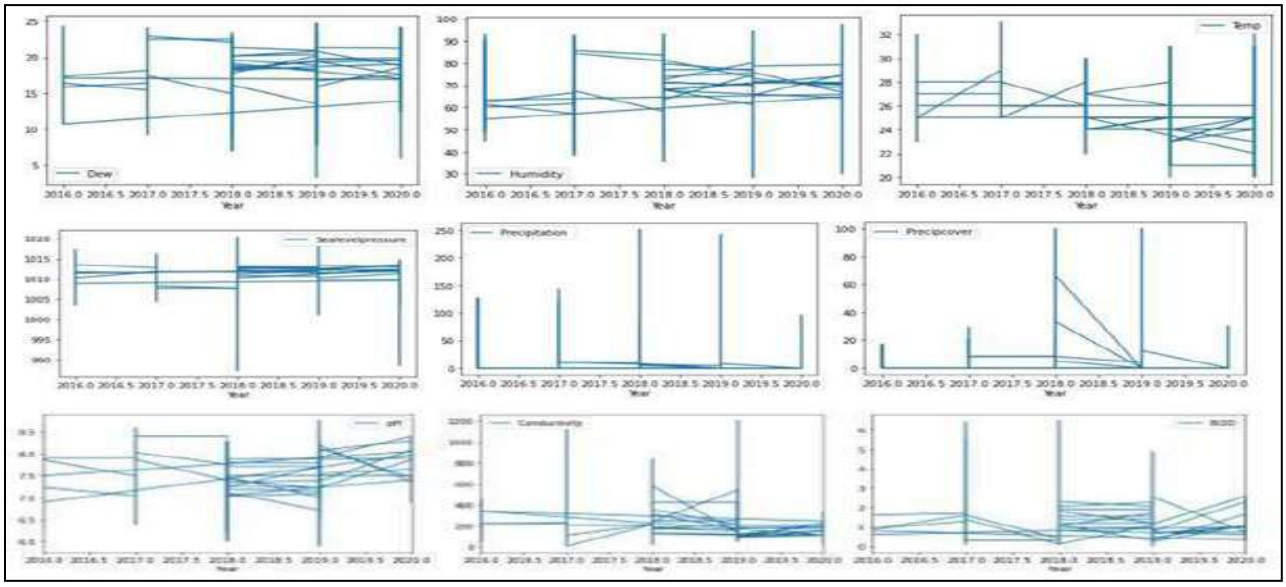

Fig.3.18. Sample Temporal Variation of Pooled Parameters of Bharathapuzha River

Incorporating the temporal variation of water quality parameters provides insights in developing the WQI prediction models. Descriptive analysis of the Bhavani River data and Bharathapuzha data are described below.

DESCRIPTIVE ANALYSIS

Descriptive analysis is a statistical technique used to summarize and describe the important characteristics of the data. Descriptive analysis of WQI parameters involves calculating various statistical measures such as mean, median, maximum, minimum, standard deviation, and range.

These measures help in identifying the range and distribution of each parameter, as well as any potential outliers or abnormal values that affect the overall WQI score. In the context of water quality assessment, descriptive analysis is used to understand the distribution and variability of different water quality parameters that contribute to the calculation of the WQI.

The descriptive statistics of the physiochemical parameters of Bhavani River are tabulated in Table IX, which illustrates each parameter along with their count, mean, maximum, minimum, and standard deviation values. The parameters included are temperature, pH, conductivity, turbidity, phenolpth alkalinity, total alkalinity, chloride, COD, TKN, ammonia, hardness, calcium hardness, magnesium hardness, sulphate, sodium, TSS, TDS, FDS, phosphate, boron, potassium, BOD, fluoride, DO, nitrate, TC, and FC. The count column shows the number of observations for each parameter, while the mean, maximum, and minimum values indicate the central tendency, upper limit, and lower limit of the values respectively. The standard deviation value reflects the spread or dispersion of the data around the mean value. Overall, the descriptive analysis provides a comprehensive overview of the physiochemical parameters with their statistical characteristics.

Table IX. Descriptive Analysis of Physiochemical Parameters of Bhavani River

Parameters	**Count**	**Mean**	**Maximum**	**Minimum**	**SDV**
Temp	10560	26	33	20	2.41
pH	10560	7.49	8.76	5.9	0.49
Conductivity	10560	161	400	6.4	127.13
Turbidity	10560	2	332	1	20.99
Phenolpth Alkalinity	10560	0	26	0	2.31
Total Alkalinity	10560	63	804	1	43.76
Chloride	10560	14	215	0	17.87
COD	10560	4	24	0.12	5.54
TKN	10560	0.1	39	0.001	1.48
Ammonia	10560	0.25	5.393	0.205	0.49
Hardness	10560	67	298	4	48.91
Ca. Hardness	10560	26	330.1	1	30.41
Mg. Hardness	10560	15	110	0.62	18.12
Sulphate	10560	6	55	0.00154	7.71
Sodium	10560	9	182	0	15.38
TSS	10560	300	300	1	134.56
TDS	10560	116	300	10	93.34
FDS	10560	125	300	0.02	100.88

Phosphate	10560	0.11	1.5	0.00063	0.19
Boron	10560	0.1	0.1	0.002	0.05
Potassium	10560	2	29	0.00845	1.57
BOD	10560	1.13	6.5	0.00036	0.9
Fluoride	10560	0.39	9.4	0	0.68
DO	10560	7.19	67	0.35	1.28
Nitrate	10560	0.54	11.423	0.0027	0.89
TC	10560	158	1800	8	220.44
FC	10560	70	1600	10	179.51

The descriptive statistics of pooled parameters of Bhavani River are tabulated in Table X which illustrates each parameter along with their count, mean, maximum, minimum, and standard deviation values the mean, maximum, minimum, and standard deviation values are given for each parameter. The parameters include physical, chemical, and biological parameters such as temperature, pH, conductivity, turbidity, alkalinity, chloride, COD, TKN, ammonia, hardness, and others. Additionally, meteorological parameters such as precipitation, wind speed, wind direction, cloud cover, and visibility are also included. The descriptive analysis is used to understand the variations and ranges of these parameters and to assess water quality and environmental conditions.

Table X. Descriptive Analysis of Pooled Parameters of Bhavani River

Parameter	**Count**	**Mean**	**Maximum**	**Minimum**	**SDV**
Temp	10560	26	33	20	2.41
pH	10560	7.49	8.76	5.9	0.49
Conductivity	10560	161	400	6.4	127.13
Turbidity	10560	2	332	1	20.99
Phenolpth Alkalinity	10560	0	26	0	2.31
Total Alkalinity	10560	63	804	1	43.76
Chloride	10560	14	215	0	17.87
COD	10560	4	24	0.12	5.54
TKN	10560	0.1	39	0.001	1.48
Ammonia	10560	0.25	5.393	0.205	0.49
Hardness	10560	67	298	4	48.91
Ca. Hardness	10560	26	330.1	1	30.41
Mg. Hardness	10560	15	110	0.62	18.12
Sulphate	10560	6	55	0.00154	7.71
Sodium	10560	9	182	0	15.38
TSS	10560	300	300	1	134.56

TDS	10560	116	300	10	93.34
FDS	10560	125	300	0.02	100.88
Phosphate	10560	0.11	1.5	0.00063	0.19
Boron	10560	0.1	0.1	0.002	0.05
Potassium	10560	2	29	0.00845	1.57
BOD	10560	1.13	6.5	0.00036	0.9
Fluoride	10560	0.39	9.4	0	0.68
DO	10560	7.19	67	0.35	1.28
Nitrate	10560	0.54	11.423	0.0027	0.89
TC	10560	158	1800	8	220.44
FC	10560	70	1600	10	179.51
Dew	10560	20.09	24.7	3.3	2.78
Humidity	10560	70.46	97.27	28.44	10.32
Sea level pressure	10560	1009.45	1020.4	987.4	2.61
Precipitation	10560	9.2	251	0	18.16
Precip cover	10560	7.04	100	0	13.42
Windspeed	10560	17.65	268.6	0.1	9.82
Wind direction	10560	154.7	337	1.2	69.26
Cloud cover	10560	48.98	99.9	1.2	18.92
Visibility	10560	5.52	10	2.2	0.98

The descriptive statistics of pooled parameters of Bharathapuzha River are tabulated in Table XI which illustrates each parameter along with their count, mean, maximum, minimum, and standard deviation values the mean, maximum, minimum, and standard deviation values are given for each parameter. The first column lists each parameter, followed by the number of observations or data points in the second column. It is evident that the temperature of the environment ranges from 24.00 to 33.00 degrees with an average of 27.61 and a standard deviation of 1.75, the environment has a relatively narrow range of temperatures, with most observations clustered around the mean value. The pH parameter represents the water's acidity or alkalinity, with a mean of 7.34 and a range of 6.72 to 8.40. Other significant parameters include Conductivity, Turbidity, Total Alkalinity, Chloride, COD, TKN, Hardness, and Sulphate, among others. Additionally, the last three parameters in the table, TC, FC, and DO, are useful indicators of water quality and are used to measure the level of microbial contamination in the water. Overall, this comprehensive data provides valuable insight into the water quality status and aids in making informed decisions towards the improvement of the water system. Another insight that drawn is that the TSS variable

has a count of 2190, which means that all observations have the same value of 300. TSS is a significant differentiating factor in the water quality environment.

Table XI. Descriptive Analysis of Pooled Parameters of Bharathapuzha River

Parameter	Count	Mean	Max	Min	SDV
Temp	2190	27.61	33	24	1.75
pH	2190	7.34	8.4	6.72	0.29
Conductivity	2190	205.51	396	86	59.38
Turbidity	2190	1.91	2	1	0.29
Total Alkalinity	2190	99.54	290	15	63.5
Chloride	2190	53.67	140	13	31.18
COD	2190	8.36	28	2	5.24
TKN	2190	0.09	0.9	0	0.06
Hardness	2190	73.07	168	30	25.66
Ca. Hardness	2190	79.49	200	16	52.12
Mg. Hardness	2190	65.36	209	8	54.21
Sulphate	2190	12.7	40.32	0	10.95
Sodium	2190	36.87	120	6	28.69
TDS	2190	205.95	500	60	111.81
FDS	2190	141.6	345	38	71.57
TSS	2190	300	300	300	0
Phosphate	2190	0.17	1.84	0.01	0.32
Boron	2190	0.1	0.1	0.1	0
Potassium	2190	5.73	41.48	1	6.34
BOD	2190	1.5	3.2	0.6	0.51
Fluoride	2190	0.33	0.58	0.1	0.1
Nitrate-N	2190	1.21	11.4	0.04	1.65
TC	2190	245.83	350	80	40.77
FC	2190	139.54	300	21	61.75
DO	2190	6.81	7.89	4.52	0.66
Dew	2190	20.09	24.7	3.3	2.78
Humidity	2190	70.46	97.27	28.44	10.32
Sea level pressure	2190	1009.45	1020.4	987.4	2.61
Precipitation	2190	9.2	251	0	18.16
Precip cover	2190	7.04	100	0	13.42
Windspeed	2190	17.65	268.6	0.1	9.82
Wind direction	2190	154.7	337	1.2	69.26
Cloud cover	2190	48.98	99.9	1.2	18.92
Visibility	2190	5.52	10	2.2	0.98

Exploratory and descriptive data analysis performed on river water quality data is extremely beneficial in understanding the characteristics of the primary data about various statistical measures. Exploratory data analysis conducted using box plots, heatmaps, pair plots, and histograms is used to discriminate between the factors of variation in water quality. The attribute distributions and correlations are investigated to find viable solutions for data preparation and data modelling requirements.

3.5. DATA PRE-PROCESSING AND DATASET PREPARATION

Data pre-processing is vital in machine learning research to ensure accurate and reliable results. Data cleaning, normalization and feature selection are three important preprocessing tasks carried out here for preparation of datasets. The main objective of data cleaning is to identify and correct errors or missing values in the data to ensure that the results of the analysis are accurate and reliable. In water quality analysis, errors arise due to various factors such as improper sample collection, measurement errors, or data entry errors. Through the process of EDA, it is revealed that certain instances of Bhavani River and Bharathapuzha River data contain missing values which required elimination. Consequently, data cleaning is performed to ensure data accuracy.

Normalization is an important step in preparing data for predictive modelling, and it is particularly relevant for water quality prediction datasets. Normalization involves scaling the values of each feature in the dataset to a common range, typically between 0 and 1. One common technique for normalization is Min-Max normalization, which involves subtracting the minimum value of each feature from all values in that feature and then dividing by the range of the feature. It is evident from the EDA that certain parameter such as conductivity, total coliform, wind speed and cloud cover of both Bhavani and Bharathapuzha river, have a wide range of observations, which requires normalization. Hence, min-max normalization is applied to standardise the values of all parameter.

Feature selection is a critical step in preparing data for predictive modelling, and it is particularly relevant for water quality datasets. Feature selection involves identifying and selecting the most relevant features from the pre-processed data which can be used in the predictive model. This help to improve the efficiency of the model and reduce the risk of over fitting, which occurs when a model is too complex and performs well on the training data but poorly on new data. In

the context of water quality prediction, relevant features include parameters such as temperature, pH, dissolved oxygen, turbidity, and other chemical and physical characteristics of the water.

In this research SelectKBest algorithm is used for feature selection. It is a widely used feature selection technique based on a statistical test, the chi-squared test, that measures the relevance of each feature to the target variable is the water quality index. The SelectKBest algorithm ranks each feature based on its score and selects the K features with the highest scores, where K is a user-defined parameter. This technique helps to reduce the dimensionality of the dataset while retaining the most important features for predicting WQI. The two features phenolpth alkalinity and boron have negative ranks for both river data and hence considered irrelevant such that they do not contribute significantly in predicting the WQI, so discarded.

Four datasets have been developed for building deep learning-based WQI prediction models. The profile of various datasets developed for this research are described below and are depicted in Table XII.

Table XII. Summary of Datasets

Dataset	Parameters	Source	Number of Instances	Number of Independent Variables	Target Variable
WQI-PCA	Physiochemical, Spatial and Temporal Parameters	Bhavani River	10560	28 Attributes	WQI
WQI-SA	Physiochemical, Seasonal, Spatial and Temporal Parameters	Bhavani River	10560	38 Attributes	WQI
WQI-BP	Physiochemical, Seasonal, Spatial and Temporal Parameters	Bharathapuzha River	2190	38 Attributes	WQI
WQI-EBP	Physiochemical, Seasonal, Spatial, Temporal Parameters Flowrate and SAR	Bharathapuzha River	2190	40 Attributes	WQI

WQI-PCA Dataset

Twenty-four physiochemical parameters, three spatial attributes, and temporal attribute of Bhavani River along with the computed WQI are included in the first dataset. A total of 10560 collected samples having 28 relevant features with WQI as the target variable forms a dataset

containing 10560 tagged instances, and this dataset is named as WQI-PCA. The sample dataset is given in Appendix A.

WQI-SA Dataset

Twenty-four physiochemical parameters, ten seasonal parameters, three spatial attributes, and temporal attribute of Bhavani River along with the computed WQI are included in the second dataset. A total of 10560 collected samples having 38 relevant features with WQI as the target variable forms a dataset containing 10560 tagged instances, and this dataset is named as WQI-SA. The sample dataset is given in Appendix A.

WQI-BP Dataset

Twenty-four physiochemical parameters, ten seasonal parameters, three spatial attributes, and temporal attribute of Bharathapuzha River along with the computed WQI are included in the third dataset. A total of 2190 collected samples having 38 relevant features with WQI as the target variable forms a dataset containing 2190 tagged instances, and this dataset is named as WQI-BP. The sample dataset is given in Appendix A.

WQI-EBP Dataset

To facilitate the implementation of the heterogenous transfer learning, the Bharathapuzha river data is extended by adding two more parameters namely flow rate and Sodium Absorption Ratio (SAR). In real time, the monitoring stations of Bhavani and Bharathapuzha river does not observe the values for these two parameters. But, recent research reports that these two parameters are important and required to be considered for building WQI prediction models.

The flow rate of a river and the SAR play pivotal roles in calculating the WQI, a crucial parameter for assessing the overall health of a water body. The flow rate influences the dilution and dispersion of pollutants, directly impacting the concentration levels of various contaminants in the water. A higher flow rate can help mitigate the adverse effects of pollutants by carrying them away and promoting better mixing. On the other hand, SAR, which evaluates the proportion of sodium to other essential ions like calcium and magnesium, is an indicator of the water's suitability for irrigation. Elevated SAR levels can indicate potential soil degradation due to sodium accumulation, leading to reduced water infiltration and plant growth. Integrating these factors into

the WQI provides a more comprehensive understanding of water quality, enabling effective management strategies to safeguard both aquatic ecosystems and human needs.

The flow rate of a river can be estimated using various methods that incorporate physical and chemical parameters. One common method involves using the Manning's equation, which relates the flow rate (Q) to the cross-sectional area (A) of the river, the hydraulic radius (R), and the Manning's roughness coefficient (n). The formula is as follows:

$$Q = (1/n) * A * R^{(2/3)} * S^{(1/2)}$$

Where Q = Flow rate (cubic meters per second), A = Cross-sectional area of the river (square meters), R = Hydraulic radius (meters), S = Slope of the river bed (dimensionless), n = Manning's roughness coefficient (dimensionless).

The Sodium Absorption Ratio (SAR) is a measure of the potential impact of sodium on soil structure and its suitability for irrigation. It is calculated based on the concentration of sodium, calcium, and magnesium ions in the water. The formula to calculate SAR is as follows:

$$SAR = (Na^{+} / \sqrt{((Ca^{2+} + Mg^{2+}) / 2)})$$

Where SAR = Sodium Absorption Ratio, Na^{+} = Concentration of sodium ions (ppm), Ca^{2+} = Concentration of calcium ions (ppm), Mg^{2+} = Concentration of magnesium ions (ppm).

Hence, the values of these two parameters for the water samples of Bharathapuzha river during the period January 2019 to December 2020 are calculated using the formulae and pooled with physiochemical and seasonal parameters. Thus, twenty-four physiochemical parameters, ten seasonal parameters, three spatial attributes, temporal attribute and two additional parameters flow rate and SAR of Bharathapuzha River along with the computed WQI are included in the extended Bharathapuzha dataset. A total of 2190 collected samples having 40 relevant features with WQI as the target variable forms a dataset containing 2190 tagged instances, and this dataset is named as WQI-EBP. The sample dataset is given in Appendix A.

3.6. TRAINING AND MODEL BUILDING

The task of predicting the water quality index is formulated as a regression problem and modelled using deep neural networks and transfer learning approaches. The WQI prediction models are built by learning the trends in the pre-processed time series data using deep learning

architectures. The training and model building for WQI prediction is carried out in four phases. First deep learning architectures such as RNN, LSTM and GRU are used as these architectures are significant in training sequence data and the WQI prediction models are developed. Next, more specialized architecture namely Temporal Fusion Transformer is employed and enhanced WQI prediction model is built by training the same dataset. In the third phase, the homogenous transfer learning technique is adopted for building the hybrid WQI prediction model. Finally, the heterogeneous transfer learning is implemented for building the generalized and a robust WQI prediction model.

Model 1: Deep Learning for WQI Prediction Models with Physiochemical Parameters

The main aim of the work is to construct an accurate WQI prediction model using physiochemical attributes and deep neural network architectures. The network such as RNN, LSTM and GRU are employed as they are designed for training sequence data. The WQI-PCA dataset is used to train the networks RNN and its variants LSTM and GRU. Various hyperparameters such as epoch, dropout, learning rate, optimizers, batch size and activation functions are defined appropriately to finetune the training and accurate WQI prediction models are developed.

Model 2: Deep Learning for WQI Prediction Models with Pooled Parameters

The objective of this work is to develop improved models for predicting the WQI by training both physiochemical and seasonal parameters using RNN and its variants. The instances of the WQI-SA dataset are given as input to the input layer of networks such as RNN, LSTM and GRU for training. The network training is done by properly setting the hyperparameters such as epoch, dropout, learning rate, optimizers, batch size and activation functions and the improved WQI prediction models are built.

Model 3: Temporal Fusion Transformer for WQI Prediction Model

The main goal of this work is to create an efficient WQI prediction model by training specialized architecture TFT with physiochemical and seasonal parameter. The adoption of TFT architecture in predicting the WQI enhances the accuracy and effectiveness of the forecasting model by leveraging the power of transformer-based time series analysis. The WQI-SA dataset is used to train the hybrid architecture with special hyperparameters such as attention windows, filter heads, value dimensions, and temporal encoder dimensions. The training of the WQI prediction

model involves the optimization of a multi-horizon forecasting objective function through the iterative updating of model parameters using backpropagation and gradient descent algorithms. Finally, an efficient WQI prediction model is constructed.

Model 4: WQI Prediction Models using LSTM Pre-trained Model

A homogenous transfer learning approach is adopted to boost the performance of WQI prediction models trained with limited data. Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. It is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned. This work uses the WQI-BP dataset and the LSTM based WQI model developed in the previous phase as pre-trained model. The knowledge gained by the pre-trained model is transferred to RNN, LSTM and GRU networks while training the WQI-BP dataset. While training and optimizing the model, the hyperparameters are correctly configured and new hybrid WQI prediction models are built.

Model 5: WQI Prediction Models using TFT Pre-trained Model

The WQI prediction models are trained using the most efficient and powerful techniques namely TFT and homogenous transfer learning to develop a hybrid model with limited data. This work uses the WQI-BP dataset, and the WQI model developed with TFT in the previous phase as the pre-trained model. The knowledge gained by the pre-trained model is transferred to RNN, LSTM, GRU and TFT networks while training the WQI-BP dataset. While training and optimizing the model, the hyperparameters are properly set and a boosted WQI models are built.

Model 6: Heterogeneous Transfer Learning for WQI Prediction Models

A heterogenous transfer learning approach is implemented to enhance the performance of WQI prediction models trained with limited data. This work uses the extended Bharathapuzha river data i.e., WQI-EBP dataset and the TFT based pre-trained model for training the RNN variants and TFT architecture using heterogeneous transfer learning. While training and optimizing the model, the hyperparameters are properly configured for each network independently and robust WQI prediction models are built.

3.7. TESTING AND EVALUATION

Testing and evaluation plays a critical role in building any prediction models using machine learning. Evaluation involves a systematic and objective examination of various aspects, such as

functionality, quality, usability, to determine their success or failure. Testing, focuses on conducting experiments to validate and verify the expected outcomes of a system. Both evaluation and testing contribute to informed decision-making, and the advancement of knowledge and innovation.

Various testing methods used in machine learning modules include holdout testing, cross-validation, bootstrapping, shuffle split, time series split, nested cross-validation, and randomized search cross-validation. In this research work, to test the model efficiency hold out testing method is used, where 80% of the total instances is used for training and the remaining 20% of the instances is for testing

EVALUATION METRICS

Evaluation metrics are essential tools for measuring the performance of machine learning models, statistical models, and other analytical methods. The metrics provide a quantitative measure of how well a model performs its intended task and enable us to compare the performance of different models. The different evaluation metrics are used in various types of data analysis and deep learning tasks. Some metrics are more appropriate for classification problems, while others are more suitable for regression. The standard metrics available in the literature that are used for evaluating the prediction models are explained variance score, mean squared error, R2 score, mean absolute error, median absolute error, root mean squared error, correlation coefficient and p value. Here, the most appropriate metrics used for evaluating the performance of WQI prediction models are MAE, MSE, RMSE and R2 score.

Mean Absolute Error

Mean Absolute Error is a widely used evaluation metric for regression models that measures the average absolute difference between the predicted and actual values of the dependent variable. The MAE is calculated by taking the absolute difference between each predicted value and the actual value, and then averaging these differences over all samples in the test set. One advantage of using MAE is that it provides a simple and intuitive interpretation of the model's performance. A lower MAE indicates that the model's predictions are closer to the actual values, while a higher MAE indicates more errors in the model's predictions. MAE is calculated using the following formulae.

$$MAE = abs(Y_a - Y_b)$$

where Y_a and Y_b are the actual responses and the predicted value, respectively, and n is the total number of variables.

Mean Squared Error

Mean squared error is a widely used evaluation metric for regression models that measures the average of the squared differences between the predicted and actual values of the dependent variable. The main advantage of MSE is that it penalizes large errors more heavily than small errors, making it particularly useful in cases where outliers or extreme values in the data have a significant impact on the performance of the model. MSE is calculated using the formulae.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_a - Y_b)^2$$

where Y_a and Y_b are the actual responses and the predicted value, respectively, and n is the total number of variables.

Root Mean Squared Error

Root Mean Squared Error is a widely used evaluation metric for regression models that measures the square root of the average of the squared differences between the predicted and actual values of the dependent variable. It is useful because it provides an interpretable measure of the average magnitude of the errors in the predicted values and penalizes large errors more heavily than small errors, making it sensitive to outliers in the data. Because RMSE is based on the same units as the dependent variable, it is more easily interpretable than MSE, which is based on squared units. Overall, RMSE is a valuable tool for evaluating the performance of regression models. It is calculated using the formulae

$$RMSE = \sqrt{(Y_a} - Y_b)^2/_n$$

where Y_a and Y_b are the actual responses and the predicted value, respectively, and n is the total number of variables.

R2 Score

The R2 score, also known as the coefficient of determination, is a commonly used evaluation metric for regression models. It provides a measure of how well the model fits the data and helps in selecting the best model for a given dataset. The R2 score ranges from 0 to 1, with 0

indicating that the model explains none of the variance in the dependent variable, and 1 indicating that the model explains all of the variance. A high R2 score indicates that the model is a good fit for the data and explains a significant portion of the variance in the dependent variable. It is calculated using the formulae.

$$R2\ score = 1 - (RSS/TSS)$$

Where, RSS is the sum of squares of residuals and TSS is the total sum of squares.

SUMMARY

The main component of this research is problem modelling and the methodology which have been well designed and have been explained in this chapter with various tasks such as data collection, exploratory data analysis, data pre-processing, dataset creation, training, testing and evaluation. The data collection, data analysis, data preprocessing and preparation of datasets have been presented in detail with sample data. The training and model building methods used in this research have been elucidated. The performance metrics used for evaluating the predictive models are also presented in this chapter. Various WQI prediction models built with the WQI-PCA dataset using DNN algorithms such as RNN, LSTM, and GRU will be presented in Chapter 4. The WQI prediction models built with WQI-SA dataset using RNN, LSTM and GRU will be discussed in Chapter 5. The WQI prediction models built with WQI-SA dataset using TFT will be elucidated in Chapter 6. The implementation of homogenous transfer learning with RNN, LSTM, GRU and TFT architecture for building WQI predictive models are explained in Chapter 7. The implementation of heterogenous transfer learning with RNN, LSTM, GRU and TFT architecture for building WQI prediction models will be discussed in Chapter 8.

Remarks

The paper titled "Exploratory Data Analysis of Bhavani River Water Quality Index Data" has been presented in International Conference on Communication and Computational Technologies - ICCCT, Jaipur, February 26-27,2022 and published in Springer Book Series. ***(Web of Science Indexed)***