Chapter 3

Identification of river water pollution level and automatic annotation of limited laboratory data

This chapter introduces a technique for detecting river water pollution levels through the estimation of the Water Quality Index (WQI) using the weighted arithmetic water quality index method. The method categorizes river water quality based on purity levels, utilizing commonly measured water parameters such as pH, turbidity, and others. Each parameter is assigned a specific weight that reflects its influence on the overall water quality, resulting in unequal contributions to the WQI. This approach enables the monitoring and management of river water contamination in real time. Additionally, the chapter discusses the automatic annotation of a limited, unlabeled laboratory dataset using the estimated WQI values.

3.1 Introduction

The investigation of river water pollution level has been a significant focus of research for an extended period. Water pollution assessment is an important part of environment monitoring as the quality of water largely affects the life of human and aquatic animals [57]. The lab-based water quality monitoring system provides ex-situ analysis of various parameters including, pH, dissolved oxygen, turbidity, *etc.* Such analysis estimates the quality of river water in terms of pollution level which helps to take preventive actions in the early stage. The estimation of water quality is also referred as Water Quality Index (WQI) [58]. It helps to identify water suitability for different purposes like drinking, irrigation, bathing, *etc.* [59]. In water pollution assessment, the laboratory sensors measure various water parameters and the generated data can be used to analyze the pollution level in the water bodies (*e.g.,* river, pond, lake, *etc.*). The pollution level in the river water can be identified if and only if the lab data is correctly annotated with the class labels (*e.g.*, excellent, good, worst, *etc.*).

The traditional techniques for river water classification do not provide the mechanism to assign labels to the dataset automatically [28,29,32]. In essence, the automatic annotation scheme is valuable in both expediting the labeling process and mitigating the errors that can arise from human involvement. In [29], the authors proposed a river classification approach based on parameters including, pH, electrical conductivity, dissolved oxygen, turbidity, *etc.* The authors use a machine learning technique to classify time-series data of the sensors attached with a boat. Randhawa *et al.* [28] developed a low-cost pollution sensor for the insight of spatio-temporal pollution pattern of river water. Support vector machine is used to train the turbidity model in the proposed system. Authors in [37] proposed an experimental platform that uses an acoustic sensor for realizing the water quality without transportation to the laboratory. In [33], the authors introduced a modelling approach for estimating pollutant distribution in river water using sensors. Next, Zhang *et al.* [32] proposed an IoT solution for flood

management and soil monitoring using wireless under ground network.

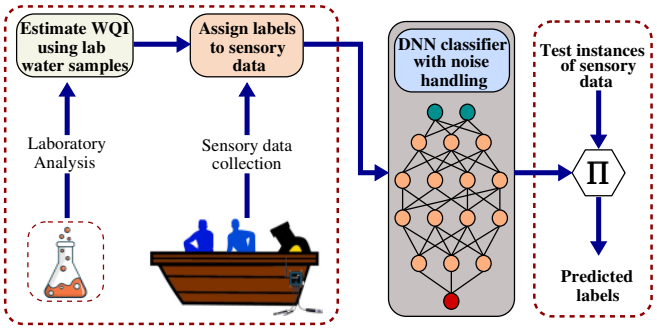


Figure 3.1: An example of river dataset collection and classification using deep leaning based classifier.

Later, the authors in [34] employed deep learning technique for retrieval of cyanobacteria pigment in river water. Finally, James rising [35] proposed an integrated assessment model that highlights the importance of river water. Fig. 3.1 illustrates an example scenario of river water dataset collection by using a boat equipped with sensors. The sensory values are sent directly to the Cloud. The water sample from the river is also collected and analyzed at the laboratory. Later, the analyzed data is also sent to the Cloud which helps in assigning labels to the unlabeled sensory data. Further, a deep learning based classifier is build to predict the labels for the testing instances.

3.1.1 Motivation of this work

Previous studies have following major limitations which motivated this work:

- The prior studies [29, 32, 34–37] on river water pollution fail to incorporate any automated annotation mechanism for labeling the laboratory data instances. In the absence of such automation, the process of annotating a laboratory dataset can be exceptionally time-consuming, stretching to months. Consequently, there exists a critical need for an automated mechanism to label a limited, unlabeled laboratory dataset.
- The previous work [28, 29, 34, 37] on river dataset use various water parameters such as pH, electrical conductivity, dissolved oxygen, turbidity, *etc.*, for assigning

labels to the data instances. Due to the impracticality of simultaneously acquiring values for all water parameters in a laboratory data instance, the annotation mechanism necessitates the use of a select subset of parameters for labeling the data instances.

In this chapter, we address the problem: *how to identify pollution level in the river water (estimation of WQI) and how to annotate automatically a limited unlabeled laboratory dataset using estimated WQI?* To address this challenge, this research introduces the creation of a mathematical model designed to estimate water quality index, which reveal the water pollution level. This model also enables the generation of labeled laboratory data using the estimated WQI. The proposed approach initiates the estimation of WQI for the purpose of assigning labels to the limited laboratory dataset. This estimation process unfolds in four sequential steps. First, it involves the selection of distinct water parameters. The second step entails the assignment of weights to these selected parameters. Subsequently, in the third step, sub-indices are computed for the chosen parameters. Finally, the fourth step leads to the derivation of the final WQI by aggregating the weights and sub-indices of the selected parameters.

3.1.2 Major Contributions

To the best of our knowledge, this is the first work to address the problem of identifying pollution level in the river water and automatically annotate the unlabeled limited lab dataset. This work makes following major contributions:

- This work presents the overarching methodology we employed for gathering the river water laboratory dataset. A predetermined route map was followed to ensure the collection of a maximum number of samples. We conducted regular visits to capture water data that exhibited spatial and temporal variations. Subsequently, all the water samples obtained were securely stored in the cloud. Data collection is of paramount importance for the analysis of fluctuating water-related parameters,

aimed at determining pollution levels. To accomplish this, we utilized glass tubes to gather data on different water parameters.

- This work involves data preprocessing to prepare it for subsequent analysis. We removed parameters with less than a 5% filled entries for that parameter, and for certain parameters, if only a small number of values were missing, we imputed them with the average of ten previous day's values.
- This work proposes an algorithm that uses k-means clustering, Pearson coefficient and CNN for selecting most distinguishable parameters from available river water parameters. We use these distinguishable parameters to estimate WQI. The estimated WQI is used for automatically assigning labels to the limited laboatory dataset such as "Bad", "Good", *etc.*.

The rest of the chapter is structured as follows. Next section describes the terminologies and notations used in this work. In Section 3.3, we discuss the mechanisms of collecting dataset from the river water. Further, Section 3.4 describes the preprocessing techniques to clean the data and make suitable for further processing. Section 3.5 proposes a mathematical approach for identifying pollution level of the river water. Next, Section 3.6 presents the experimental analysis of proposed approach. Finally, Section 3.7 concludes the chapter.

3.2 Preliminaries and problem statement

In this section, we first describe the different terminologies used in this work. Later, this section covers a brief description of the problem associated with pollution level identification in river water using lab data (estimation of WQI) and the overview of solution.

3.2.1 Preliminary

The pollution level of the river can be identified by analyzing the different parameters of the water including pH, Dissolved Oxygen (DO), nitrates, Biochemical Oxygen Demand (BOD), Fecal Coliform (FC), *etc.* The analysis of parameters incorporates two methods, namely, lab based and sensor based. In the lab based method, the water sample collected from the river is analyzed in the laboratory to identify the pollution level of the river. The lab based method incorporates the estimation of WQI that uses sub-indices and calculate weights for each parameter. The lab based analysis gives the result of the pollution level after a specific time (time needed for evaluation), which sometimes even a month. This delay in pollution level prediction hampers the real-time monitoring of the river.

Definition 3.1 (Parameter) *A characteristic or quantity that can influence the output or behavior of a particular system is defined as a parameter.*

Definition 3.2 (Sub-indexing) *Sub-indexing is a technique that converts the actual values of different parameters to a common scale (dimensionless number) by using a mathematical relationship between actual and scaled values.*

Definition 3.3 (Water quality index) *Water Quality Index (WQI) is a unique rating that describes the water quality to signify its suitability for drinking, irrigation, bathing, etc. The estimation of WQI incorporates various parameters, such as pH, dissolved oxygen, biochemical oxygen demand, and so on.*

3.2.2 Problem statement and overview of solution

Pollution level identification in the river water helps in protecting the river water for sustaining the survival of human beings on the Earth, as discussed in the introduction. The laboratory data instances collected from the river water are unlabeled, therefore

assigning appropriate labels to the dataset is a tedious task. This work, therefore, addresses the problem of automatically annotating lab data instances by using estimated WQI.

Overview of the solution: This work proposes a mathematical model that etsimates WQI and automatic annotation of unlabeled dataset using estimated WQI. The estimation of WQI comprises four steps: selection of distinct parameters, assigning weights to the chosen parameters, calculating sub-indices for the selected parameters and aggregating weights and sub-indices to estimate final WQI. Later, the estimated WQI is used for assigning labels to the lab dataset automatically.

3.3 River water dataset collection

In this section, we cover the process of dataset collection utilized in our research work. We specially covers the creation of two different datasets, namely, lab-based dataset and sensory dataset. We prepared the lab data set by collecting the data analysed in the lab and the sensor data by using the sensors. Also, it covers the overall approach and procedure utilized to collect water bodies data, incorporating five rivers, namely, Ganges, Hindon, Godavari, Yamuna, Brahmaputra, and Bangalore lakes. Next, The collected data is then transmitted to the base station or network server using a wireless communication protocol such as BLE, Wi-Fi *etc.*. Therefore, we discuss various wireless communcation techniques. We mainly cover different topics incorporated in the river water data collection in the following sections:

3.3.1 Methodology for collection of river water dataset

In this section, we discuss the overall methodology we adopted to collect the river water dataset. The data collection process includes a multi-parameter sensor on a boat to collect the river water data at different times such as morning, afternoon and evevning of the day on a prepared route map to collect the maximum samples. We

executed a regular boat rides to collect spatially and temporally varying water data using attached sensors. Next, we stored all the water samples collected from the sensors to the cloud. Data collection is crucial for analyzing the varying concentrations of water-related parameters to figure out the pollution level. We used non-stationary mobile sensors for collecting the different water parameters. Fig. 3.2 shows the detailed approach for data collection followed by our team. This enabled us to catch the river

Figure 3.2: Step-by-step procedure followed for river watre data collection.

water pollution using various parameters in different seasons such as winter, rainy and summer and at different locations across the river and map the water pollution as a function of time and space, resulting in a water pollution system. The river water pollution system can catch and guess water contamination and advice in determining the strength of sanitation interference and their detailed causes, helps in certainly handling the control of diseases in populations that live in the river valleys, recognize time-changing sources of contamination such as inadequate sewage treatment plants, and bring about conciousness of water quality and pollution issues.

The primary purpose of this dataset is to find out the pollution of the water due to anthropological activities of people living in the surroundings of rivers, industrial water drainage, sewage, agricultural wastes, *etc.*. Following the investgation of the geo-tagged spatial and temporal water quality data, we apply physics-based and analytical techniques to recognize the enlargement of pollution, introduction of interpolation technique for sparse data, and perform inverse analysis to find sources of contamination. The dataset which we have collected is present at [1]. This work incorporates two kinds of river water datasets, *i.e.,* the lab dataset created off the site and the other sensory

dataset collected on-site. In the next sections the description of lab and sensor datasets along with all parameters are given in detail. The sample name, sample date, time, month name, river, location,type, lattitute, longitude are common parameters in lab data set and sensor data set.

3.3.2 Wireless communication protocols for IoT applications

This section explains how to choose the right wireless protocol among various wireless communication protocols for IoT applications required to transmit the river pollution data from a river to the nearest gateway or network server. The IoT is percolating its way into water pollution application. There has been an increase in the demand and usage of wireless protocols. As a result, there are a variety of wireless technologies that are available for use in IoT applications. The various factors to consider while choosing a wireless technology are: bandwidth, communication range, power consumption and cost. While making a choice of a wireless protocol, there is a trade-off between bandwidth, communication range, power consumption and cost. While the bandwidth and communication range are decided based on the application, the cost and power consumption must be minimized as much as possible. If you want to send a large amounts of data, you can make use of Wi-Fi tecchnology as Wi-Fi offers high bandwidth and a good range for data transmission. On the downside, Wi-Fi modules consume relatively high power and are expensive. If you want data transmission over a short range, you can make use of technologies like Bluetooth or RFID. RFID or radio frequency identification as name suggests makes use of radio frequency wave for communictaion. This is used for very short-range communications. Generally, Bluetooth technology is used For short-range communications in meters. Bluetooth Low energy is the best choice for the applications that do not require continuous connection but depend on high battery life. Zigbee technology is the best for long range data transmission and if you want to compromise on the data rate as it is a low-power, low data rate protocol. LoRa is the

best fit If we need a very long range of data transmission for small amounts of data because it provides a low-data rate and long range communication with low power. After collection of the geo-tagged data in the csv files, there is the need of data cleansing by removing the noise. After prepocessing the raw data, converted into a standard scale and then used by Deep Learning algorithm to decide the pollution level of river water. The pollution level is transmitted to the gateway using long-range communication technology. The available communication protocols are Long Range Wireless Area Network (LoRaWAN), Bluetooth Low Energy (BLE) and WiFi *etc.*. The description of communication protocols utilized in determining the pollution level of river water are as given in Table 3.1.

Table 3.1: Different Wireless Communication Protocols for water pollution monitoring

Technology	Topology	Frequency	Coverage (line-of-sight)	Data Rate	Power	Cost
RFID	P2P, Star	Varies	3 m	Varies	Low	Low
Bluetooth	P2P, Mesh, Star	2.4 GHz	100 m	1-3 Mbps	Low	Low
WiFi	Star	5 GHz	70m	1Gbps	High	Low
Zigbee	Mesh, Star	2.4 GHz	200m	250 kbps	Low	Moderate
LoRaWAN	Star	865 MHz	10 km	300 Kbps	Low	Low

In our research work, we have several low cost sensors used for collecting water sensory data produces vast sensory data. These sensors work on battery and the lifespan of around a couple of years with the same battery. They are able to transmit information to gateway somewhere between $1km$ to $10km$ away. The data further transmitted to the gateway or network server using an appropriate communication protocol. We used LoRa which is a radio frequency transmission method which allows the devices to consume a very little amount of power and transmit informtation over a long distance. LoRa is operating in Industrial, Scientific and Medical free open frequency bands because of that we are transmitting information to more than $5kms$ without license.

3.3.3 Off-site (ex-situ) laboratory dataset

Here, in this section, we discuss the lab dataset in detail. The complete procedure for Lab data collection and analysis includes recognition of interest-based sites on the condition of water, an arrangement for water sample collection in the test tube. Next, developing a route map across the river assuring highest potential data acquisition. Further, analyzing the various parameters of sample water in the laboratory. The water sample collected from the identified site is sent to the laboratory for detecting and analyzing values of the various parameters like DO, pH, nitrates, *etc.* The estimated value with tagged GPS coordinate is stored in a data file upon which preprocessing techniques are applied to curtail down the noisy contents. Next, cleansing the data by preprocessing for creating data visualizations such as heat-maps and transmitting the raw data along with heat-maps to the open-source digital platform such as Cloud for storage as shown in Figure 3.3. Afterwards, the data is stored, which can be used for estimating river pollution using learning models such as, deep learning model discussed in the next chapter. The lab dataset consists of 561 records. Each

Figure 3.3: Steps involve in Lab dataset collection of the river water.

record incorporates general information such as Sample_name, Date, Month_number, Month_name, River_name, Location_name, Type_water; type of water such as normal, industry- mixed, and domestic-mixed *etc.*, Lattitude, Longitude, Sampling_time attributes of water data along with the various parameters as shown in Table 3.2. It contains readings of significant cations (+ ions) found in natural water such as calcium, magnesium, sodium and potassium, whereas major anions (- ions) include chloride, sulfate, carbonate, bicarbonate, fluoride and nitrate. Along with the significant ions,

heavy metals are the inorganic elements that affect human health by causing several diseases though their levels are generally very minute. Next, the lab dataset incorporates Chemical Oxygen Demand (COD), the quantity of oxygen used for chemical oxidation of organic waste to inorganic end products. It measures the oxidizable organic contamination in water. Next, the BOD_5 is a measure of the amount of oxygen that bacteria will consume under aerobic conditions while decomposing organic wastes.To balance these organic wastes the minimum DO level should be 2-7 mg/L then only the BOD can be estimated in Laboratory condition at fixed temperature *i.e.* $20°C$ for 5 days.

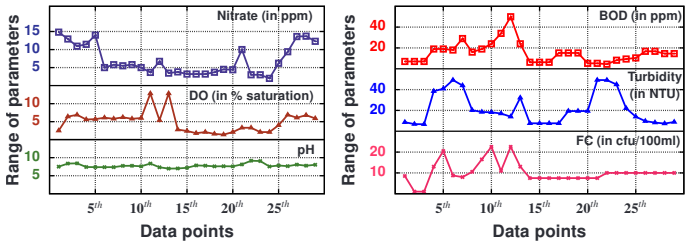
Moreover, Fecal coliform (FC) in the waste water mainly results from human or animal excreta. However, the waste from paper and pulp mills, plant materials and agricultural runoff can also increase fecal coliform concentration in waste water. Further, Common nitrite (NO_2) sources include acid rains, agricultural runoff. Chloride comes from erosion, weathering of rocks, industrial wastes and domestic wastes. In addition, total hardness refflects the amount of dissolved calcium and magnesium salt in water resulting from weathering of rocks and anthropogenic activities. Total suspended solids tells about the suspended particles can come from soil erosion, runoff, discharges, stirred bottom sediments or algal blooms. Sodium and Potassium come from erosion, weathering of rocks, industrial wastes and domestic wastes. Magnesium and Calcium come from erosion, weathering of rocks, industrial wastes and domestic wastes. The lab dataset incorporates the readings for more than 20 heavy metals such as Lithium, Beryllium, Aluminum, Vanadium, Chromium, Manganese, Iron, Cobalt, Nickel, Copper, Zinc, Arsenic, Selenium, Rubidium, Strontium, Cadmium, Cesium, Barium, Thallium, Lead, and Uranium are analyzed. Drinking water with massive amounts of metal can cause metal poisoning leading to death. Next, Radioactive material like Technetium is responsible for cancer. Additionally, there are seven water parameters: pH, DO, EC, temperature, turbidity, Ammonium and nitrate concentra-

tion of point samples are also measured in the labratory for validation purposes. pH parameter tells whether the water is acidic or basic and Turbidity parameter gives information about how clear the water is while Pressure parameter tells about the pressure of water. Dissolved oxygen parameter in the water is very essential for aquatic animals. Conductivity plays a role for passing the light through water. Nitrate (NO_3)and nitrite (NO_2) parameters are very important.

Table 3.2: Lab dataset parameters along with its unit collected in our Lab dataset.$\#NTU$, $\#psi$, $\#ppm$, $\#cfu$, AND $\#rfu$ denote Nephelometric Turbidity unit, Pounds per square inch, parts per million or milligrams per liter (mg/L), Colony forming unit, and relative fluorescence unit, respectively

Parameter(Acronym/Unit)	Parameter(Acronym/Unit)	Parameter(Acronym/Unit)
potential of Hydrogen (pH)	Turbidity (NTU)	Pressure (psi)
Total Solids (ppm)	Dissolved Oxygen (ppm)	Conductivity (μS/cm)
Chemical Oxygen Demand	Biochemical Oxygen Demand	Total Suspended Solids
Fecal Coliform (cfu_100ml)	Total Coliform (cfu_100ml)	Nitrate (NO_3) (ppm)
Nitrite (NO_2)	Chloride (ppm)	Total Hardness
Total Dissolved Solids (TDS)	Temperature ($^\circ$C)	Tryptophan (rfu)
Chlorophyll_a (rfu)	Colored Dissolved Organic Matter (CDOM) (rfu)	Ammonia
Magnesium (Mg)(ppm)	Thorium (TH) (ppm)	Fluoride (ppm)
Sodium (Na)(ppm)	Potassium (K) (ppm)	Calcium (Ca) (ppm)
Nickel (60Ni_1)(ppm)	Uranium metal (238U_2)	Lithium metal (7Li_2)
Beryllium metal (9Be_2)	Aluminium metal (27Al_1)	Vanadium metal (51V_1)
Chromium metal (53Cr_1)	Manganese metal (55Mn_1)	Iron metal(56Fe_1)
Cobalt metal (59Co_1)	Copper metal(63Cu_1)	Zinc metal(66Zn_1)
Arsenic metal(75As_1)	Selenium metal (82Se_1)	Rubudium metal (85Rb_2)
Strontium metal (88Sr_2)	Cadmium metal (111Cd_2)	Cesium metal (133 Cs_2)
Barium metal (137Ba_2)	Thalium metal (205Tl_2)	Lead metal (208Pb_2)

The parameters selected in this work (from lab dataset) has shown significant variation in their value over a month. Part(a) and Part(b) of Fig. 3.4 illustrates the variation in the parameters of water sample collected from Ganges on different days in a specific month. It can be observed from Fig. 3.4, the parameter pH is nearly constant through out the month whereas, other parameters shows significant variation in their values. Thus, we have unlabelled lab dataset. We construct a mathematical model to estimate



(a) pH, DO, and Nitrate (b) FC, turbidity, and BOD.

Figure 3.4: Variation in values of different parameters of lab data collected from Ganges in Varanasi.

WQI and annotate the lab data automatically using estimated WQI. Further, we use labelled lab data to annotate the sensory data automatically. In the next section, we discuss the sensory dataset in detail.

3.3.4 On-site (in-situ) real and massive sensory dataset

Here, we discuss the sensory dataset in detail along with its collection procedure. Similar to the lab-dataset, the sensory dataset collection needs a mobile sensors attached to the boat. The complete procedure for sensory data collection and analysis includes recognition of interest-based sites on the condition of water, an arrangement of a boat and attaching an automated sensor for monitoring spatial distribution of diverse parameters of water data. Next, developing a route map across the river assuring highest potential data acquisition. Further, execution of daily sensor-equipped boat rides to acquire temporally and spatially changing data. Finally, data collection by downloading it from all sensors, cleansing the data by preprocessing and transmitting the raw data to the open-source digital platform such as Cloud for storage. The steps involve in the sensory dataset collection is illustrated in Fig. 3.5. The interested site detection covers

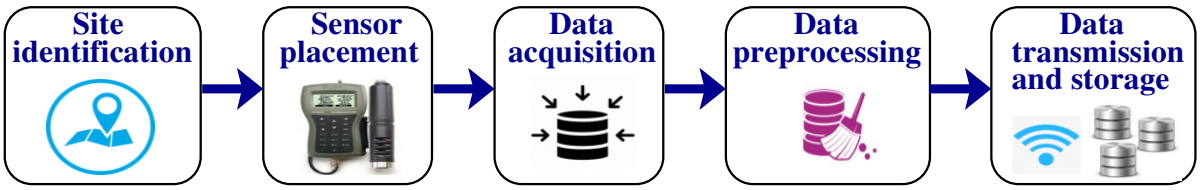


Figure 3.5: Steps involve in sensory dataset collection of the river water.

choosing the areas where the channels are having waste of industry, agriculture, and anthropogenic activities mixing with the river water. Hanna multiparameter HI-9829, a submersible automated sensor is placed on a boat for acquiring the dataset. The regular boat rides are performed on a regular interval such that a large dataset can be collected. The variation of different water parameters are ananlyzed by performing boat rides with variable days and seasons in the year. This time stamped and GPS

tagged data is than preprocess by using different filtering and segmentation technique and then transferred to the cloud when it can be permanently stored.

The sensory dataset incorporates different water pollution parameters. The agricultural and urban runoff along with treated wastewater from municipal and industrial outlets impacts the water quality. The water pollution is measured by several factors; including physical (such as temperature, turbidity, color, taste and odour of water *etc.*), chemical (such as pH, Dissolved Oxygen, Total Solids, biological oxygen demand, metals *etc.*) and biological parameters (such as faecal coliform, total coliform, tryptophan *etc.*). Additionally, The chemical characteristics of natural water are a reflection of the soils and rocks with which the water has been in contact. Parameters such as temperature, electrical conductivity, pH, dissolved oxygen and turbidity are measured using thermometric, electrometric and turbidometric techniques. These parameters are measured in-situ using the real-time data logging sensors. Microbial and chemical transformations affect the chemical characteristics of water which depend on inorganic and organic compounds. Inorganic compounds may dissociate to varying degrees, to cations and anions. In the sensory dataset, more than 3,26,000 records are available. Each record incorporates general information such as sample_id, water_body_container, water_body_location, sample_date, sample_time, Lattitude and Longitude characteristics of water data along with other in-situ measured water data parameters as shown in Table 3.3. The other sensory water data parameters, namely, pH, DO, Temperature, Turbidity, EC, Ammonia concentration, Nitrate concentration, CDOM, Chlorophyll-a, Tryptophan *etc.* are measured during each boat ride. The pH exceeding the range (such as $6.5 - 8.5$) may damage the outer part of adult fishes, namely, eyes and gills. The sustainability of aquatic life depends on the proper oxygen content in the dissolved form, where a fast decrease in the DO represents upraised organic pollution in the water. The temperature tells about the temperature of the river water surface. Next, the turbidity indicates the suspended particulates in the water. The more the total

suspended solids in the water more the turbidity. EC gives information about the total suspended solids and ions (cations and anions) available in the water. The higher the total suspended solids in the water, the higher the EC and the lower the ions lower the EC. Therefore, low EC water can cause ions deficiency in the human as drinking water is the primary source of ions.

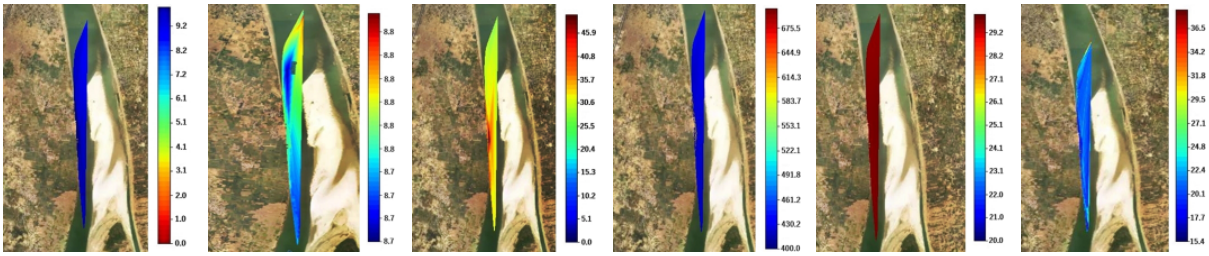
Further, Ammonia can enter the water via direct means such as municipal effluent discharges and the excretion of nitrogenous wastes from animals, and indirect means such as nitrogen fixation, air deposition, and runoff from agricultural lands. In addition, Ammonia occurs in water bodies due to the microbiological decomposition of nitrogenous compounds in organic matter. Fish and other aquatic organisms excrete ammonia. It may be directly discharged into water bodies by some industrial processes or as a component of domestic sewage or animal slurry. Next, sewage, industrial wastes and agricultural run-off are the critical sources of Nitrates in water. Moreover, Colored or chromophoric dissolved organic matter consist of naturally occurring water-soluble, biogenic, heterogeneous organic yellow substances essential for the biochemical cycle in river water. Organic chemicals cause disagreeable tastes and odours in drinking water, resulting in water contamination. They handle the light penetration by absorbing Ultra Violet light, nutrient availability and ecosystem productivity. High levels of organic content due to Natural or synthetic processes prohibit phytoplankton's growth, causing decreased DO amount in the water. Next, Chlorophyll-a is the photosynthetic pigment that causes the green color in algae and plants. It absorbs sunlight and converts it to sugar during photosynthesis. Chlorophyll-a concentrations are an indicator of phytoplankton abundance. Further, Tryptophan is an alpha - amino acid that is used in the biosynthesis of proteins.

Additionally, our research work incorporates the use of in-house algorithms to run on the sensory data to create data visualizations such as heat-maps and transmitting the raw data along with heat-maps to the open-source digital platform such as Cloud

Table 3.3: Sensory dataset parameters along with its unit collected in our sensory dataset. $\#NTU$, $\#psi$, $\#ppm$, $\#cfu$, $\#\mu S/cm$,AND $\#rfu$ denote Nephelometric Turbidity unit, Pounds per square inch, parts per million or milligrams per liter (mg/L), Colony forming unit, microSiemens per centimeter, and relative fluorescence unit, respectively

Parameter	Unit	Parameter	Unit
potential of Hydrogen (pH)	-	Turbidity	NTU
Pressure	psi	Temperature	° C
Dissolved Oxygen (DO)	ppm	Electrical Conductivity (EC)	$\mu S/cm$
Total Dissolved Solids (TDS)	mg/L	Nickel (Ni)	ppm
Chlorophyll_a (chl_a)	RFU	Colored Dissolved Organic Matter (CDOM)	RFU
Ammonia concentration	ppm	Nitrate concentration	ppm
Total Dissolved Solids	-	Tryptophan (tryp)	RFU
Nickel	ppm	Ammonia	ppm

for storage. The heat map is like a table or cross table visualizations without numerical values. The heatmaps help in getting the clear cut view of the situation much better as color gradient variation is more well-liked for humans instead of sizable, technical tables of data, which hampers the logical ability of a non-technical person. Heat maps have ability to derive valuable insights from vast sensory data. Data visualizations help in determining the relationships among various parameters with each other, which assists in giving glimpses how the river system works in different dynamic situations. In heat maps, the color of the cell relates the value. Spatio-temporal pollution observations are derived and represented via color-based heatmaps. The large sensory dataset collected from different five rivers show low to high variation in the values of different parameters recorded on a different time in a single day. Significant variations are observed when



(a) DO. (b) pH. (c) Turbidity. (d) EC. (e) Temperature. (f) Tryptophan.

Figure 3.6: Heat maps of different parameters in the sensory dataset [1], on a specific date (for river Ganges in Varanasi).

the sensory data are collected in different month of a year. Parts (a) to (f) of Fig. 3.6, illustrates the heatmaps of Dissolved Oxygen (DO), Potential of Hydrogen (pH), Turbidity, Electrical conductivity (EC), Temperature, and Tryptophan of river Ganges in Varanasi on a specific date. In the next section, we elaborate on the multi-parameter sensor used in our work.

3.3.5 Hanna multi-parameter HI-9829 sensor used in our work

Here, in this section, we discuss about the sensor incorporated in our research work. We have used Hanna multi-parameter HI-9829 meter having multiple sensors capable enough to provide high resolution, geo-tagged, in-situ sensing of multiple parameters via a moving multi-parameter sensor platform to collect the massive sensory data as shown in Fig. 3.7. The HI-9829 multi-parameter portable meter with GPS is the perfect tool for monitoring multiple water parameters. It is the best for river water to measure up to 14 water quality parameters including pH, conductivity, TDS, Temperature, Turbidity, Dissolved oxygen, Chloride, Nitrates, Ammonia *etc.*. Next, it offers an optional GPS module making it easy to track measurement locations and return to them in the future. Further, it can log up to 44000 records and offers an optional autonomous logging probe that has a logging memory of up to 140000 individual samples and 35000 complete samples. The HI-9829 offers a water-to-cloud feature, where its autonomous logging probe connects to the cloud, enabling direct transfer of water data to the cloud for storage and further processing. It is having application software to manage locked data and map sample locations.

3.4 Data preprocessing techniques

In this section, we cover various steps utilized in our research work to preprocess the river water data to make it suitable for applying a deep learning approach to assess river water pollution using automated annotation. Data preprocessing is an important step

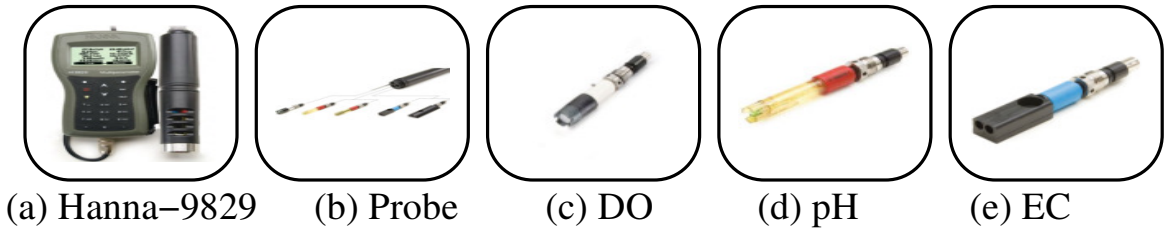


Figure 3.7: Portable waterproof Hanna multi-parameter HI-9829 meter

in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task. Initially, after collecting water data, data is preprocessed using various techniques in the sequence to select the water parameters. In our work, the data preprocessing approach comprise two steps as follows, Step 1: Availability of parameters, Step 2: Selection of distinct parameters. In the first step, while assessing parameter availability, we took into account unfilled entries. In the second step, in the process of selecting distinct parameters, our approach begins with a k-means clustering method. Subsequently, we employ Pearson correlation to identify the least correlated parameters. Finally, we utilize Convolutional Neural Networks (CNN) to detect the distinct parameters and arrive at the definitive set of parameters.

Fig. 3.8 illustrates the first phase involved in the proposed approach. Our proposed methodology consists of four stages: estimating WQI, automatically annotating sensory data, constructing a deep learning classifier, and predicting class labels for new test data instances. This chapter focuses on presenting the initial phase. The first phase revolves around generating labeled lab data. Initially, the input is unlabeled lab data. Further, we estimate the WQI (water pollution level) and allocate the labels to the lab dataset. Estimating the WQI involves four key steps: selecting water parameters, determining weights for these parameters, computing sub-indices for the selected parameters, and ultimately aggregating the weights and sub-indices of these parameters. In the next

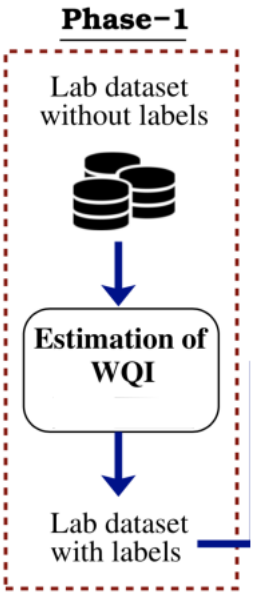


Figure 3.8: Illustration of first phase involve in the proposed approach.

section, we delve into the process of selecting unique parameters.

3.4.1 Selection of distinguishable parameters

In this section, we address the diverse steps involved in our work for the selection of distinctive water quality parameters. The selection of the parameters (*e.g.*, temperature, pH, electrical conductivity, turbidity, *etc.*) for assessment of the river water pollution is an essential step in the estimation of WQI. Let $R = \{r_1, r_2, \cdots, r_m\}$ be the set of m rivers whose water quality is to be estimated on a set $p = \{p_1, p_2, \cdots, p_l\}$, of l parameters. $\mathbf{D}^L_{R*p*T}$ denotes the dataset analyzed in the lab, where $R * p * T$ is a tensor. The value of T is an integer representation of actual system time during data collection. Mathematically, selection of parameters for estimating WQI can be represented as

$$\mathbf{D}^L_{R*p*T} \xrightarrow[selection]{parameter} \mathbf{D}^L_{R*p'*T}, \tag{3.1}$$

where, p' ($p' \subseteq p$) is the reduced set of parameters for estimating WQI. This dimension reduction of parameters p' helps in faster and convenient training of a classification

model on the dataset. In this work, the selection of the parameters incorporates following sequential steps.

3.4.1.1 Availability of parameters:

The first and foremost step in the selection of parameters is to verify its availability. Let x_{ij} is an element of dataset $\mathbf{D}^{L}_{R*p*T}$, $\forall\ i \in N$ and $j \in p$. The availability of a parameter $p_j \in p$ is verified by an iterative observation of all element x_{ij} in $\mathbf{D}^{L}_{R*p*T}$. Let μ_j denotes % unfilled entry of a column j in the dataset. The condition for availability of parameter j is given as $\left(\frac{\mu_j}{N} \times 100\right) > \phi$, where ϕ is the threshold determined experimentally for selecting a parameter.

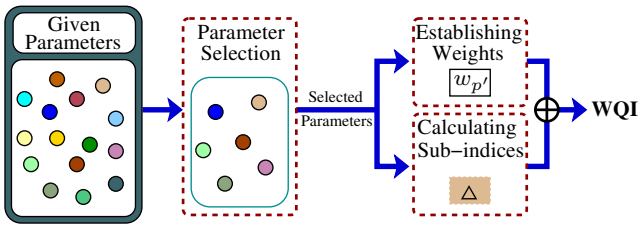


Figure 3.9: Block diagram for estimation of water quality index.

All the parameters that satisfy the condition are available parameters. The unfilled entries in the dataset may arise due to infrequent monitoring of a parameter with respect to other. For example, the value of parameter fluoride is less frequently observed in comparison to that of parameter pH. Therefore, we can ignore the value of parameter fluoride as its observations are less. Some parameters satisfied the criteria to become available parameters. Still, for some available parameters, some of the values were missing, we filled those entries using the previous value because, in one day, there will be no drastic change in parameter values in general. In this paper, the river water labdataset has 61 parameters discussed in the previous chapter. These 61 parameters are reduced to 17 after verifying availability of all the parameters at a threshold of $\mu = 27\%$. These parameters are further passed to next step for reducing dimension based on the distinguishable chracteristics and low correlation value.

3.4.1.2 Selection of distinct parameters:

Next, we select the distinct parameters from the available parameters. The selection procedure incorporates clustering of dataset into k clusters using k-means clustering technique. The k-means clustering is a unsupervised learning algorithm. If the input data does not contain any target or dependent variable, then k-means clustering is the best fit for that situation. In our research work, we try to make groups or clusters of data instances based on the different values for the distinct parameters of lab dataset. Afterward, each cluster is assigned with an arbitrary class label such that it will acts as an phantom label before the actual labels are assigned to the given dataset. The Pearson correlation [60] is estimated after k-mean clustering to fetch out least correlated parameters. Further, a CNN model is incorporated to estimate error for all least correlated parameters for calculating final set of reduced parameters p'. A CNN model can capture the spatial dependencies of the data and if the location or place of the same data changes, still CNN will be able to classify properly. The CNN model works in two steps: feature extraction and Classification. Feature Extraction is a phase where various filters and layers are applied to the data to extract the information and features out of it and once it's done it is passed on to the next phase i.e Classification where they are classified based on the target variable or label of the problem. It consists of four layers such as input layer, convolution layer having activation function, pooling layer and fully connected layer. The input layer gets our lab dataset as input. It is required to normalize the input between 0 to 1 before passing it to the model. Next, the convolution layer is responsible for extracting or detecting the features by applying filters many times to create a feature map. Once we get the feature map, an activation function such as ReLU is applied to it for introducing nonlinearity. The pooling layer is applied after the Convolutional layer and is used to reduce the dimensions of the feature map which helps in preserving the important information of the input data and reduces the computation time. Using pooling, a lower resolution version of input is

created that still contains the large or important elements of the input image. Till now we have performed the Feature Extraction steps, now comes the Classification part. The Fully connected layer is used for classifying the input data into a label. This layer connects the information extracted from the previous steps (i.e Convolution layer and Pooling layers) to the output layer and eventually classifies the input into the desired label. The softmax activation function is applied on the output layer to classify the data.

Procedure 3.1 illustrates all the step involved in the selection of parameters. In the river dataset after applying this procedure we obtain 6 distinguishable and least correlated parameters, *i.e,* DO, pH, BOD, FC, NO_3, and turbidity.

3.5 Estimation of river water pollution level (WQI)

In this section, first of all, we discuss a method for estimating the WQI of different rivers. During the estimation of WQI, we transform and aggregate the selected water quality parameters into a dimensionless number using weights assigned to the parameters. There are following ways to assign the weights to the parameters; a) equal weights are assigned if the parameters have similar importance and b) unequal weights are assigned if some parameters are less or more important than others. Different parameters for assessment of river water quality have a different range of values. It is thus required to transform these parameters into sub-indices (scaling values of parameters to a common scale). These sub-indices and obtained weights are aggregated to obtain a final index WQI. Fig. 3.9 illustrates complete steps involved in the estimation of WQI. The estimation method for WQI consists of following four significant steps; 1) Selection of distinguishable parameters 2) Establishing weights, 3) Calculating the sub-indices, and 4) Aggregation of weights and sub-indices.

Procedure 3.1: Selection of parameters

Input: Dataset $\mathbf{D}^{L}_{R*p*T}$ with N instances and p parameters, k (class count);
Output: Set of reduced parameters p' ($p' \subseteq p$);
/*Variable initialization*/
$\max_{iteration} \leftarrow 1000$, iter $\leftarrow 1$.
/*Random labels assignment*/
Randomly select k centroids $c_1, c_2, \cdots c_k$.
Verify parameters availability.
while *iter* $\leq \max_{iteration}$ **do**
  for $i \leftarrow 1$ *to* N **do**
    /* n(.) is cardinality of set*/
    for $j \leftarrow 1$ *to* $n(p)$ **do**
      /* Estimating nearest neighbor*/
      $\forall\ x_{ij} \in \mathbf{D}$, nearest centroid $\in \{c_1, c_2, \cdots, c_k\}$.
      Assign x_{ij} to the obtained cluster.
  for $q \leftarrow 1$ *to* k **do**
    Update centroids $(c_1, c_2, \cdots, c_k)$ with mean value.
  iter $\leftarrow$ *iter* $+ 1$;
Assign arbitrary label to all the k clusters.
Obtain dataset $\mathbf{D}'$ with these arbitrary labels.
Estimate powerset $\mathcal{P}$ for p (set of parameters).
$\mathcal{P}' \leftarrow$ *Pearson_correlation*($\mathcal{P}$)
/* Feature selection using CNN */
for $s \in \mathcal{P}'$ **do**
  Calculate error e_s on s using CNN with arbitrary labels.
  if e_s *reaches limiting value or* 0 **then**
    return s

Function *Pearson_correlation* ($\mathcal{P}$)
  begin
    for $s \in \mathcal{P}$
      if $|s| \neq 0$ or 1 **then**
        Calculate $\rho_{\mathbf{a},\mathbf{b}} = \frac{cov(\mathbf{a},\mathbf{b})}{\sigma_{\mathbf{a}}\sigma_{\mathbf{b}}}$, **a,b** $\in s$.
        Select the parameters that are least correlated.
      Assign selected parameters in $\mathcal{P}$.
    return $\mathcal{P}$.
  end

3.5.1 Selection of recognizable parameters

In the previous section, we have seen various sequential steps to choose the final set of water parameters. Initially, we had more than sixty water parameters in the lab dataset. First, we compared the filled values of all parameters with the threshold value. Next, we removed those parameters unsatisfied with the threshold condition for determining the water pollution level. After this comparison, still, we had more than fifteen parameters left with us. Further, we used the k-means clustering technique to divide the dataset into k different clusters by assigning phantom labels. After clustering, we used the Pearson's correlation coefficient to decide the least correlated parameters. Next, we incorporated a CNN model to estimate the error for all least correlated parameters and determine the final set of parameters to find the water pollution level. Finally, we discussed the algorithm used for selection of distinct water parameters utilized to determine the water pollution.

3.5.2 Establishing weights for selected parameters

The parameters selected in the previous section have a different influence on the estimation of river water quality. This influence of one parameter over another, managed by assigning equal or unequal weights to different parameters [61]. The assignment of equal weights to different parameters indicate similar importance of them. However, it is impractical to assign equal weights to all the parameters as different parameters impart diverse influence on estimating the water quality. If the permissible limit of a parameter used for obtaining different classes (e.g., water supply, irrigation, drinking and so on) is low, then the minor variations in the value of the parameter can influence the water quality to a large extent. Therefore, the weights assigned to these parameters are higher.

There are mainly two methods in the literature [62] to assign weights to the different parameters. The first method relies on the expert opinion where water quality analysts

suggest the weights to be assigned. Due to the varying opinion of diverse expertise, this method proves to be less relevant for the accurate estimation. The second method incorporates different permissible limits and guidelines for estimating weights. These methods are widely used in the existing literature for assigning the weights. In this work, we are using a weight assignment technique that incorporates weights used by the National Sanitation Foundation (NSF) [2] against various water parameters.

Parameters	Weights
Dissolved Oxygen	0.17
Potential of Hydrogen	0.11
Biochemical Oxygen Demand	0.11
Fecal Coliform	0.16
Nitrates	0.10
Turbidity	0.08
Temperature Change	0.10
Total phosphate	0.10
Total Solids	0.07

(a) NSF

Reconstruction

Parameters	Weights
Dissolved Oxygen	0.23
Potential of Hydrogen	0.15
Biochemical Oxygen Demand	0.15
Fecal Coliform	0.22
Nitrates	0.14
Turbidity	0.11

(b) Proposed

Figure 3.10: Weight establishment in proposed approach using NSF [2].

Let b is the set of parameters used by NSF for estimating the weights, $b = \{$DO, pH, BOD, FC, NO_3, turbidity, TC, TP, TDS$\}$. The set of selected parameters p' from the Section 3.4.1 is $p' = \{$DO, pH, BOD, FC, NO_3, turbidity$\}$. The weight of parameters in p' is estimated with the help given weights in part (a) of Fig. 3.10 taken from [2]. The weight w_{p_j} for parameter $p_j \in p'$ is calculated as follows

$$w_{p_j} = \frac{w_{b_j}}{\sum_{b_j \in p'} w_{b_j}}, \tag{3.2}$$

where, w_{b_j} is the weight assigned to j^{th} parameters in b, $\forall b_j \in p'$. The resultant weight after applying Eq. 3.2 is illustrated in part (b) of Fig. 3.10.

3.5.3 Calculating the sub-indices for selected parameters

Next, the parameters selected in the Section 3.4.1 are having different dimension with a wide range of values. Therefore, it is required to transform these value to a standard

scale for appropriate analysis of the water quality. For example, the measurement unit of nitrates (NO_3) is ppm, and its values lie in a range of $0-100$, whereas, the range of pH is $2-12$. Similarly, turbidity measured in nephelometric turbidity unit in the range $0-100$ and dissolved oxygen varies from $2-140$ measured in % saturation. Therefore, it is necessary to rescale or standardize the value of the parameters before performing the aggregation with weights to generate WQI. The resultant output after the transformation to a standard scale is commonly known as sub-indices.

The six parameters considered in this work are transformed on a common scale using the formula discussed in [63]. Each parameter for a data element x_{ij} ($\forall i \in N \text{and} j \in n(p')$) is transformed as follows:

(a) Dissolved Oxygen: The value of DO (% saturation) in range $2-140$ can be transformed on a standard scale using following mathematical formulation

$$\begin{aligned}
&\textit{if } (x_{ij} < 50) \textit{ then } x_{ij} \leftarrow e^{-(x_{ij}-98.33)/36.067},\\
&\textit{else if } (50 \leq x_{ij} < 100) \textit{ then } x_{ij} \leftarrow (x_{ij} - 107.58)/14.7,\\
&\textit{else } x_{ij} \leftarrow (x_{ij} - 79.543)/19.054.
\end{aligned}$$

(b) Potential of Hydrogen: Next, the value of pH shows different trend in comparison to that of other parameters (DO, FC, *etc.*) and varies in a range of $2-12$. The transformation in pH value can be obtained as follows

$$\begin{aligned}
&\textit{if } (x_{ij} = 7) \textit{ then } x_{ij} \leftarrow 1,\\
&\textit{else if } (x_{ij} > 7) \textit{ then } x_{ij} \leftarrow e^{(x_{ij}-7)/1.082},\\
&\textit{else } x_{ij} \leftarrow (7 - x_{ij})/1.082.
\end{aligned}$$

(c) Biochemical Oxygen Demand: The BOD varies in range of $0-30$ and is measured in ppm units. The scaling formulation for BOD is as follows

$$\begin{aligned}
&\textit{if } (x_{ij} < 2) \textit{ then } x_{ij} \leftarrow 1,\\
&\textit{else } x_{ij} \leftarrow x_{ij}/1.5.
\end{aligned}$$

(d) Fecal Coliform: Further, the bacterial FC is measured in unit of colonies/ml and its values may reach beyond 100000 in the water bodies. The transformation of FC is

given as

$$\begin{aligned}
&\textit{if } (x_{ij} < 50) \textit{ then } x_{ij} \leftarrow 1,\\
&\textit{else if } (50 \leq x_{ij} < 5000) \textit{ then } x_{ij} \leftarrow (x_{ij}/50)^{0.3010},\\
&\textit{else if } (5000 \leq x_{ij} < 15000) \textit{ then}\\
&\quad x_{ij} \leftarrow (x_{ij}/50 - 50)/16.071,\\
&\textit{else } x_{ij} \leftarrow (x_{ij}/15000) + 16.
\end{aligned}$$

(e) Nitrates: The ppm concentration of the NO_3 varies from $0 - 100$ and can be transformed to a common scale as follows

$$\begin{aligned}
&\textit{if } (x_{ij} \leq 20) \textit{ then } x_{ij} \leftarrow 1,\\
&\textit{else if } (20 < x_{ij} \leq 50) \textit{ then } x_{ij} \leftarrow e^{(x_{ij}-145.16)/76.28},\\
&\textit{else } x_{ij} \leftarrow (x_{ij}/65).
\end{aligned}$$

(f) Turbidity: Finally, the nephelometric turbidity unit is used to measure the turbidity of the water. The turbidity varies in range $0 - 500$ and scaled on common scale as

$$\begin{aligned}
&\textit{if } (x_{ij} \leq 5) \textit{ then } x_{ij} \leftarrow 1,\\
&\textit{else if } (5 \leq x_{ij} < 10) \textit{ then } x_{ij} \leftarrow x_{ij}/5,\\
&\textit{else } x_{ij} \leftarrow (x_{ij} + 43.9)/34.5.
\end{aligned}$$

The approach discussed in [2] proposes a similar transformation technique using the Q-value by incorporating rating curve. Upon estimating the scaled value form the curve fitting, we got a similar result to that are discussed in the aforementioned mathematical formulation of parameters transformation.

The permissible limit of all the parameters are used by index developers to obtain various classes (*e.g.,* water supply, irrigation, drinking and so on) depending upon the water quality. The permissible limit for different parameters can be obtained from [64]. In this work we are using linear interpolation technique [65] to obtain sub-indices, if the actual parameter lie in between two classes. Suppose, the increment in the value of FC harm the fish life in a river but the simultaneous increment in DO, jointly indicate the

quality of water is suitable for fish life. This contradiction affecting the water quality estimation resolved by using Eq. 4.0 and Eq. 4.1.

Let x_j^u and x_j^l be the upper and lower bound of the permissible limit for the parameters in p'. x_u and x_l denotes the sub index value of upper and lower bound. The sub-index x_{ij} for element $x_{ij}, \forall i \in N$ and $j \in n(p')$ is calculated as:

$$x_{ij} = x_l - \left[(x_l - x_u) \left(\frac{x_{ij} - x_j^l}{x_j^u - x_j^l} \right) \right]. \tag{3.3}$$

$$x_{ij} = x_l - \left[(x_l - x_u) \left(\frac{x_j^l - x_{ij}}{x_j^l - x_j^u} \right) \right]. \tag{3.4}$$

Eq. 4.0 is used for the case when the quality of water decreases with increases in the value of the parameter (*e.g.,* FC), whereas, Eq. 4.1 can accommodate the case when the increase in the parameter increases the quality of water (*e.g.,* DO).

3.5.4 Aggregation of weights and sub-indices

Finally, the estimated weights and the calculated sub-indices are aggregated to obtain the value of the WQI. This work uses a weighted mean for aggregating the weights and sub-indices. The weighted mean resembles with the additive or the arithmetic operation for estimation of WQI [61]. The estimated weights for the selected parameters are multiplied with sub-indices and combine to estimate WQI. The six selected parameters with their weights and sub-indices are aggregated as per following two mathematical expressions

1. If all the parameters are equally important (same weights are assigned), then

$$WQI = \sum_{i=1}^{N} \sum_{j=1}^{n(p')} x_{ij}, \tag{3.5}$$

where, $n(p')$ is the cardinality of set p' ($n(p') = 6$).

2. If the parameters are less or more important (unequal weights are assigned), then

$$WQI = \sum_{i=1}^{N} \sum_{j=1}^{n(p')} w_{p'_j} x_{ij}. \tag{3.6}$$

The obtained WQI is a single-valued dimensionless number against the multiple parameters used for analyzing the water quality. Thus WQI can be considered as a function of water influencing parameters that play a vital role in water quality assessment of a given source (water bodies like river, lake, *etc*). This work uses a scale of 100 inspired by the scale used by [2] that summarizes results from distinct parameters of Lab Data. Table 3.4 summaries the range of WQI and its corresponding class labels. It also illustrates the integer labels used in this work against the value of WQI.

Table 3.4: Range of WQI for different class labels and their corresponding integer representation used in this paper.

WQI	$80-100$	$70-80$	$60-70$	$50-60$	$30-50$	$0-30$
Class label	Excellent	Very Good	Good	Medium	Bad	Very Bad
Integer label	6	5	4	3	2	1

3.6 Experiments and results

In this section, we first discuss the implementation details of the proposed approach. Later, the approach is evaluated, under different parameters settings on the river dataset [1].

3.6.1 Implementation details

This section discusses the implementation of the proposed approach. We implemented the proposed approach in Python programming language using PyTorch library. The implementation of the first phase is as follows.

1. The implementation of the first phase incorporates a separate function for selecting an appropriate number of parameters. These parameters further used for

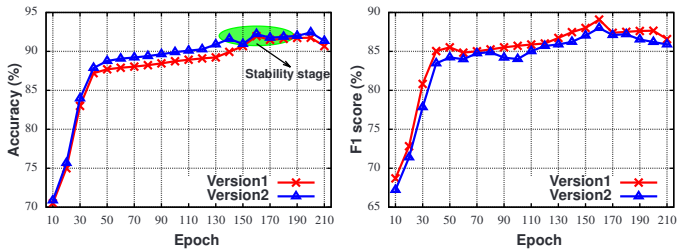
estimating WQI using python code.

2. Next, we implemented a separate function to annotate the lab data automatically using estimated WQI.

3.6.2 Experimental results

In this section, several experiments are carried out to evaluate the performance of the proposed approach and provide answers to the following questions:

- What is the appropriate epoch count for achieving the best performance of the proposed approach? (Section 3.6.3)
- How does the performance of the proposed approach influence with a different combinations of parameters for labels assignment? (Section 3.6.4)
- What is the class-wise accuracy of the proposed approach when selected parameters are used? (Section 3.6.5)



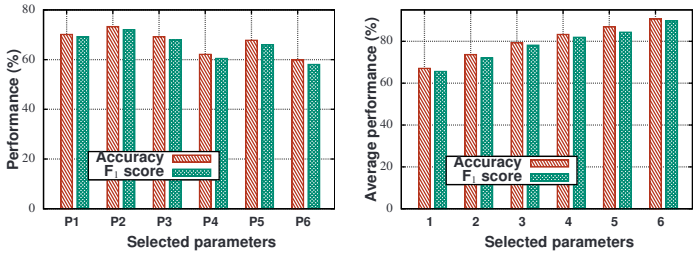
(a) Accuracy (b) F1-score.

Figure 3.11: Performance measure of river datasets during training.

3.6.3 Number of epochs

The experimental evaluation starts with the estimation of a suitable number of epochs for training the model. The performance of the proposed approach on the training data is determined using two versions; namely, Version1 and Version2. In Version1, we have used the 6 selected parameters, whereas, in Version2, all 17 available parameters are used. Part (a) of Fig. 3.11 illustrates that the accuracy at the training time increases

rapidly up to 150 epochs and then a nominal improvement is observed. The proposed approach achieves maximal accuracy at 160 epochs (≈ 92%). Similar to that of accuracy, F1-score shows impartial behaviour on variations in epochs count as illustrated in part (b) of Fig. 3.11. Thus, the experimental analysis from next section onwards is performed on 160 epochs.



(a) Single parameter (b) Combination of parameters.

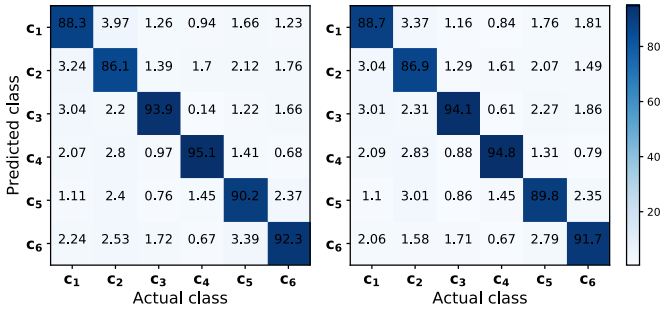
Figure 3.12: Impact of the selected parameter(s) on proposed approach performance.

3.6.4 Number of river parameters

Further, this work studies the impact of parameters which are selected (in Section 3.4.1.2) for estimating WQI, on the performance of the classifier. Let, **P1**, **P2**, **P3**, **P4**, **P5**, and **P6** denotes the selected parameters DO, pH, BOD, FC, nitrates, and turbidity, respectively. Part (a) of Fig. 3.12 illustrates the impact on the classifier performance when the labels of lab dataset are estimated using a single parameter, and then sensory labels are assigned. The pH values outperform to that of others and achieve accuracy and F1-score of around 76% and 75%, respectively. The combination of these parameters enhances the accuracy of the built classifier. The maximal accuracy is observed when all six parameters are used.

Part(b) of Fig 3.12 illustrates the average accuracy of the different combinations of parameters. As the parameters combinations increase, the accuracy showed continuous improvement and achieved maxima for the combination of all 6 parameters. This improvement in accuracy is a matter of the fact that the increase in the number of

parameters improves the distinguishable features learn by the classifier. The continuous increment in the accuracy reaches a saturation stage after specific number of parameters. In this work, this saturation point is achieved after application of 6 selected parameters.



(a) (b)

Figure 3.13: Class-wise accuracy. Part(a) for selected 6 parameters and Part(b) for all 17 parameters.

3.6.5 Class-wise accuracy

Later, this work estimates the class-wise accuracy of the proposed approach on river dataset at selected 6 and available 17 parameters. The effect of parameters is a coalition of both increment and decrement in the class-wise accuracy when parameters increase from selected 6 to available 17. Fig. 3.13 illustrates that the accuracy of classes $\mathbf{c_1}, \mathbf{c_2},$ and $\mathbf{c_3}$ shows a minor improvement when all the 17 parameters are used on the other hand remaining 3 classes ($\mathbf{c_4}, \mathbf{c_5},$ and $\mathbf{c_6}$) suffers minimal accuracy compromise. The coalition in the accuracy is because of the classes $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$ have a sufficient number of instances which are uniquely identified with the increase in the parameters count. In the case of classes $\mathbf{c_4}$, $\mathbf{c_5}$, and $\mathbf{c_6}$ the instances are lower in number to that upper three classes; therefore, the increment in the parameters count hamper the performance of the classifier.

3.7 Conclusion

Initially, this chapter presented the detailed description of data collection process. The data collection process covered methodology for data collection in detail. After data collection, it is transmitted to the nearest processing unit or network server. Therefore, next this chapter covered the wireless communication protocols. In our reserach work, we have utilized two datasets. Therefore, it provided the information about lab dataset and sensory dataset along with its parameters in depth. Next, this chapter covered the Hanna multi-parameter sensor used to collect the sensory data. Further, this chapter presented data preprocessing techniques to select the appropriate parameters to identify the water pollution level. The estimation of water pollution level consisted of four steps and obtained labeled lab dataset. The experimental evaluation showed that the suitable number of epochs is 160 to achieve the maximal accuracy during training by considering two groups of parameters, one consisting of 6 selected parameters and another with 17 available parameters. Next, the experimental evaluation also showed that the maximum accuracy is achieved for all 6 selected parameters compared to a single parameter. Finally, the class-wise accuracy have a minimal accuracy compromise for selected 6 parameters compared to 17 available parameters which is approximately($\approx 1\%$).