

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
MOHAMED BOUDIAF UNIVERSITY - M'SILA

FACULTY OF TECHNOLOGY
DEPARTMENT OF ELECTRONICS
N° :



FIELD: ELECTRONICS
OPTION: INSTRUMENTATION

**Dissertation submitted in partial fulfillment of the requirements
for the Academic Master's Degree**

Presented by:
BAKACHE Habiba & FOUDILI Kaouther

Title:

**Development of an Intelligent Tool for
Water Quality Assessment**

Defended on June 10th, 2026 in front of the jury:

BNNACER Hamza	Professor – Univ. of M'sila	Chair
DJERIOUI Mohamed	Professor – Univ. of M'sila	Examiner
LADJAL Mohamed	Professor – Univ. of M'sila	Supervisor
OUALI Mohammed Assam	Professor – Univ. of M'sila	Co-Supervisor

Academic year: 2025 / 2026

Acknowledgements

First and foremost, all praise is to Allah, the Almighty, for granting us the strength and perseverance to complete this work.

We would like to express our deepest gratitude and sincere thanks to our esteemed supervisors, **Prof. Mohamed Ladjal** and **Prof. Mohammed Assam Ouali**, for their invaluable guidance, insightful advice, and patience throughout the duration of this research.

Without forgetting the jury members: **Prof. Bennacer Hamza** and **Prof. Djerioui Mohamed**, who agreed to examine and evaluate this modest work.

We also extend our profound appreciation to the Head of the Department, **Dr. Mustapha Tabbakh**, for his kind treatment, warm welcome, and immense support, as he spared no effort in guiding and assisting us throughout our path.

Our special thanks and recognition also go to **Dr. Khennouf salah** for their generous academic support and constant guidance, which have been a true pillar for us during our academic journey.

Furthermore, we would like to express our profound appreciation to those who guided us during our field internships:

(Internship at Sonatrach): we extend our sincere thanks to my supervisor, **Mr. Guendaz**, for his excellent supervision and continuous support throughout the internship at Sonatrach.

(Internship at Sonelgaz): we extend our heartfelt gratitude to the supervisor, **Mr. Saci Cherif**, for his invaluable guidance and dedication during the internship at Sonelgaz.

Finally, our thanks go to all the respected professors who assisted us directly or indirectly, and to everyone who extended a helping hand or a word of encouragement throughout our studies.

Thank you all



﴿وَاٰخِرُ دَعْوَاهُمْ اَنْ الْحَمْدُ لِلّٰهِ الْعَالَمِيْنَ﴾

Dedication

To the sources of endless giving and unconditional love... my dear father and beloved mother, you are my bedrock, my guiding light, and my constant strength. Your wisdom, continuous guidance, and prayers in the dark of night have turned every hardship into ease. To you, I proudly dedicate the fruit of this effort and success. May Allah grant you long, healthy, and blessed lives.

To my wonderful brothers and sisters, my lifelong companions, Thank you for your unwavering presence, your constant encouragement, and for easing my path every single day so that I never wanted for anything throughout this academic journey.

To my dear uncle, Thank you for your unconditional support from the very beginning, for always standing firmly by my side, and for being a true source of strength and motivation.

To my closest friends and colleagues, Thank you for being my rock during the challenges we faced, which we overcame together with resilience, and for sharing every moment of perseverance and ambition on these university benches



﴿وَأَخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ الْعَالَمِينَ﴾

Dedication

To the sources of endless giving and support... my dear father and beloved mother,
May Allah grant them long and healthy lives. You are my bedrock, and through
your prayers, all hardships were overcome. To you, I dedicate the fruit of this
effort.

To my wonderful family, and especially my relatives living here,
Heartfelt thanks for your constant support and presence by my side, and for
easing the way for me so that I never wanted for anything throughout my journey.

To my dear and beloved aunt,

Who has been like a sister and a true friend to me. Thank you for your
unconditional support in everything, and for standing by me from the very beginning
of my university studies until this very moment.

To my dearest friends,
Thank you for your sincere presence and uninterrupted support. Thank you for
being each other's rock during the ordeal we faced, which we overcame together
with strength and resilience.

To my fellow classmates and colleagues,
Who shared the ups and downs of this academic journey with me, and with whom I
shared moments of hard work and perseverance

Foudili kaouther

Table of contents

Summary	I
List of Abbreviations.....	IV
List of Figures	V
Liste des Tableaux.....	VI
Introduction	Erreur ! Signet non défini.

CHAPITRE I: Water Quality Index

I.1 Introduction.....	4
I.2 History of the Concept of Water Quality Assessment.....	4
I.3 Different Uses of Water	6
I.4 Definition of Water Quality Index	6
I.4.1 Categories of WQI	6
I.4.1.1 Public Indices	6
I.4.1.2 Specific Consumption Indices.....	6
I.4.1.3 Designing/Planning Indices.....	7
I.4.1.4 Statistical Indices	7
I.4.2 Methodological Framework for WQI Calculation	7
I.4.2.1 Water Quality Measurement Parameter Selection	7
I.4.2.2 Standardized Scale Transformation of Raw Data Parameters	7
I.4.2.3 Providing weights to the parameters	7
I.4.2.4 Combining Sub-Indices Using a Mathematical Expression.....	7
I.4.3 Different methods for WQI determination	8
I.4.3.1 National Sanitation Foundation Water Quality Index (NSFWQI).....	8
I.4.3.2 Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI)	9
I.4.3.3 Oregon Water Quality Index (OWQI)	10
I.4.3.4 Weighted Arithmetic Water Quality Index Method (WAWQI)	10
I.5 Selection of Water Quality Parameters	12
I.5.1 Water quality standards.....	12
I.5.2 Water quality measurement parameters.....	13
I.5.2.1 Physicochemical parameters.....	13
I.5.2.1.1 pH Level	13
I.5.2.1.2 Temperature.....	13
I.5.2.1.3 Total Dissolved Solids (TDS)	13

I.5.2.1.4 Electrical Conductivity (EC)	13
I.5.2.1.5 Dissolved oxygen	13
I.5.2.2 Pollution parameters	14
I.5.2.2.1 Sulfate...	14
I.5.2.2.2 Nitrate Concentration	14
I.5.2.2.3 Chloride Concentration.....	14
I.6 Water Quality Index (WQI)	14
I.7 Justification for use of WQI.....	16
I.8 Conclusion	16

CHAPTER II :Machines learning methods

II.1 Introduction.....	17
II.2 Study Area Description.....	17
II.2.1 Geographic Location.....	17
II.2.2 Climate Characteristics	17
II.2.3 Water Source Description	18
II.3 Data Sources and Sensors Information.....	18
II.4 Data Preprocessing Strategy	18
II.4.1 Data Cleaning	19
II.4.2 Data Normalization and Standardization.....	19
II.4.3 Correlation Analysis of Water Quality Parameters	20
II.5 Machine Learning Framework.....	21
II.5.1 Support Vector Regression (SVR).....	21
II.5.2 Random Forest Regression (RFR).....	23
II.6 Feature Importance and selection Strategy	24
II.6.1 Importance for Instrumentation and Machine Learning.....	25
II.7 Model Training and Validation	25
II.7.1 Training Process.....	25
II.7.2 Cross-Validation Method	26
II.7.3 Testing and Evaluation Metrics	27
II.8 Conclusion.....	28

CHAPTER III : Results and discussions

III.1 Introduction	29
III.2 Data Collection and Sampling:.....	29
III.2.1 Sampling Strategy.....	29

III.2.2 Water Quality Index Calculation	31
III.3 Proposed framework	34
III.4 Results and Analysis	35
III.4.1 Model Selection	35
III.4.2 Model Evaluation Metrics.....	36
III.4.3 Training and Testing results.....	36
III.4.4 Feature selection and importance analysis using RFR model.....	40
III.5 Discussion and Practical Implications	43
III.5.1 Intelligent Data Aggregation	43
III.5.2 Robustness and Generalization	44
III.5.3 Practical Real-Time Deployment	44
III.6 Conclusion:.....	44
Conclusion.....	46
Bibliography	49

List of Abbreviations

AI : Artificial Intelligence
ANN : Artificial Neural Networks
BOD / BOD5 : Biochemical Oxygen Demand (5 days)
CCME-WQI : Canadian Council of Ministers of the Environment Water Quality Index
Cl⁻ : Chloride
COD : Chemical Oxygen Demand
CRNN : Convolutional Recurrent Neural Network
DO : Dissolved Oxygen
EC : Electrical Conductivity
EQI : Environmental Quality Index
KFCV : K-Fold Cross-Validation
MAE : Mean Absolute Error
ML : Machine Learning
MSE : Mean Squared Error
NO₃⁻ : Nitrate
nse: normalized num of excursions
NSFWQI : National Sanitation Foundation Water Quality Index
NTU : Nephelometric Turbidity Unit
OWQI : Oregon Water Quality Index
pH : Potential of Hydrogen
PO₄³⁻ : Phosphate
R² : Coefficient of Determination
ReLU : Rectified Linear Unit
RF : Random Forest
RFR : Random Forest Regression
RMSE : Root Mean Square Error
SO₄²⁻ : Sulfate
SRDD : System of Radionuclide Detection and Description
SS/TSS : Suspended Solids / Total Suspended Solids
SVM : Support Vector Machines
SVR : Support Vector Regression
T° / Temp : Temperature
TDS : Total Dissolved Solids
Turb : Turbidity
WA : Weighted Arithmetic
WAWQI : Weighted Arithmetic Water Quality Index
WHO : World Health Organization
WQI : Water Quality Index
WQM : Water Quality Monitoring

List of Figures

Figure 1.1. Historical Development of the WQI Model.....	5
Figure 1.2. Phases of WQI Development.....	8
Figure 2.1. Geographic Location of Tilesdit Dam in Bouira, Algeria	17
Figure 2.2. Mathematical Formulation and Visual Distribution of Normalization versus Standardization Methods.....	20
Figure 2.3. SVR Predictive Function Components: (a) Soft Margin, (b) varepsilon-Insensitive Loss Function, (c) Optimal Hyperplane and Support Vector.....	22
Figure 2.4. Workflow of Random Forest Regression and Feature Importance Ranking.....	24
Figure 3.1. Evolution of Descriptive Parameters from 2016 to 2018	30
Figure 3.2. Variation in Sample Features of Six Selected WQ Parameters	33
Figure 3.3. Flowchart of the Proposed Methodology for WQI Prediction Using ML Model.	34
Figure 3.4. WQI Values Prediction Results for SVR and RFR Models in the Training Phase...	39
Figure 3.5. Scatter Plot of Predicted WQI Values Using SVR and RFR Models in the Training Phase.....	39
Figure 3.6. WQI Values Prediction Results for SVR and RFR Models in the Testing Phase	39
Figure 3.7. Scatter Plot of Predicted WQI Values Using SVR and RFR Models in the Testing Phase.....	40
Figure 3.8. Importance of Input Parameters (RF)	41

List of tables

Table 1.1. Water Quality Rating as Per Different Water Quality Index Methods	11
Table 1.2. Standard Maximum Values of Certain Parameters to be used in this study	12
Table 1.3. Classification of Water Quality Based on the WQI Scale and Corresponding Water Status	16
Table 2.1. Mathematical Formulation of Performance Evaluation Metrics	28
Table 3.1. Descriptive statistics for used water quality variables	30
Table 3.2. Standard Values and Relative Weights of Water Quality Parameters.....	31
Table 3.3. Relative weights and standard values of water quality variables used in WQI calculation.....	32
Table 3.4. Weight of Physicochemical Parameters and Algerian Standard for Surface Water Quality (2016 and 2018)	33
Table 3.5. An Example of the Calculation of the Water Quality Index (WQI).....	33
Table 3.6. Average Performance Metrics of RF and SVR Methods for WQI Prediction Using 5-Fold Cross-Validation on Training and Testing Datasets.....	38
Table 3.7. Performance Metrics for FR-Based Feature Selection and WQI Prediction.....	42
Table 3.8. Performance Metrics for RR-Based Feature Selection and SVR-Based WQI Prediction	42

Introduction

One of the most basic needs for existence is water, which also has a significant impact on human society. For instance, water is crucial for productivity in the agricultural sector and all facets of socioeconomic growth. Water quality is essential to human health, ecological health, and a number of economic activities. Poor water quality can have far-reaching consequences that affect several facets of society and the environment. Water source degradation is the term used to describe the deterioration of freshwater resources due to a variety of circumstances. Affects the availability and quality of water from natural sources such as reservoirs, rivers, lakes, and aquifers. Degradation of water sources is caused by a variety of human activities and environmental conditions. In order to reduce environmental issues and maintain the wellbeing of ecosystems, human populations, and economies, water quality is crucial [1].

Therefore, maintaining biodiversity, safeguarding drinking water, and putting into practice sustainable water policy all depend on continuous monitoring. In order to simplify data for stakeholders, water quality assessment often entails testing a variety of physicochemical indicators, including dissolved oxygen (DO), pH, conductivity, TDS, temperature, and turbidity, and getting the results into a WQI. The basic classification of WQI into categories like "excellent" to "unsuitable" [2].

The most common source of consumers is surface water. In this way, in order to make the surface water drinkable, stricter treatments would be needed. Studying surface water bodies is important and necessary. Surface water is typically used for irrigation and drinking. In order to assess if the surface water is suitable for drinking and other uses, we investigated its physicochemical characteristics.

The Water Quality Index (WQI) of water can indicate whether the general quality of water bodies poses a risk to several uses of water, including drinking water supply, recreation and aesthetics, agriculture and animal irrigation water, and aquatic life habitat. WQI is a single-value indication of the water quality. It incorporates the data pool that is produced following the calculation of weights for the used parameters [3]. Indices like: CCME-WQI, NSFQI, OWQI, and WAWQI are the most often used WQI. These WQI have been useful for estimating the quality of water in certain locations. The many quantities and kinds of water quality parameters are frequently the basis for these indices. The water quality (WQ) parameter is contrasted with the corresponding regional standards. The widely utilized WAWQI in poor nations was used to discuss the WAWQI approach. Based on their acquired numerical value, the WQI condenses vast amounts of water quality data into straightforward

phrases, such as excellent, good, bad, extremely poor, and unsuitable. Multiple WQ parameters are combined into a single mathematical formula by the WAWQI.

The WAWQI is chosen to determine the WQ of a certain location for consumption. The researchers can use it with ease and simplicity. WQI's ability to condense complex data into understandable language is one of its advantages. Based on several WQ factors, the WQI gives a single figure that represents overall WQ at a specific location and time. The public can solve a number of water-related issues by using the WQI as a tool to compare the WQ of different water sources. For this reason, we have selected this WAWQI for water quality resilience [4].

The traditional techniques of calculating water quality indices are characterized by great technical and economic difficulties due to the need for periodical sampling and continuous laboratory analyses, which are time consuming and expensive. They also are susceptible to human error or data gaps due to sensor and monitoring equipment failures. In this present era, a revolutionary transformation is happening by combining Machine Learning (ML) methods with environmental measurement and sensing systems to overcome these obstacles. Machine learning algorithms (Support Vector Regression (SVR) and Random Forest (RF) are characterized by an excellent ability to handle large data streams, derive complex patterns, and perform non-linear modeling of the association between several physicochemical variables without the need of pre-specified mathematical equations.

The main problem addressed by this research revolves around the efficiency and capability of machine learning algorithms (SVR and Random Forest) in building intelligent predictive models capable of accurately predicting the WQI. We did also feature selection of parameters using RF and found the most influential parameter in WQI prediction.

The work carried out consists of three chapters, presented as follows:

Chapter one provides a detailed literature review and historical perspective of water quality assessment. It focuses on the emergence and growth of Water Quality Indices (WQIs) as analytical tools since the middle of the twentieth century. This chapter emphasizes the importance of these indices in the reduction of complex multi-dimensional physicochemical parameters into single simplified dimensionless index scores. This synthesis allows decision makers and non-technical stakeholders to quickly assess the state of water resources and understand the trends of environmental degradation without delving in the sophisticated laboratory details.

Chapter Two is devoted to materials and methods, elaborating the engineering methodology used for data preparation and the theoretical basis of artificial intelligence. This chapter discusses data pre-processing mechanisms, including data cleaning, such as linear interpolation to handle missing values due to sensor failures, statistical evaluation techniques, and data normalization to ensure stable training dynamics. It also provides a structural and architectural analysis of the intelligent algorithms used, namely Random Forest (RF) and Support Vector Regression (SVR) as well as the statistical metrics used to assess the performance of the models.

Finally, Chapter Three represents the empirical core and field study of this research, focusing on water quality assessment and prediction at the Tilesdit Dam in the Bouira province as a case study. This chapter presents the analytical calculation and intelligent prediction results of the WAWQI. This is followed by the display and discussion of the predictive performance of the machine learning models SVR and RF in estimating the WAWQI values. The chapter concludes by discussing the engineering significance of deploying these low-power intelligent models into embedded systems within the dam's local monitoring station to provide real-time alerts for the sustainable protection of this vital water resource.

The last part of this study presents a general conclusion revisiting the different steps performed, summarizing the main results obtained and emphasizing the future perspectives envisioned for the development of intelligent ecological monitoring systems.

Chapter I

Water Quality Index

I.1 Introduction

Water quality is a dynamic attribute influenced by various physical, chemical, and biological parameters. Accurate assessment therefore requires an understanding of the interrelationships among these parameters and the external conditions affecting them. Individual parameters alone are insufficient for evaluating overall water quality, which creates the need for effective reporting tools. Water quality indices (WQIs) offer a strategic approach for synthesizing complex data into a single value that is understandable to stakeholders. Despite their usefulness in summarizing water quality information, WQIs have several limitations, including possible data loss through aggregation, subjective parameter weighting, and the absence of a universal index.

I.2 History of the Concept of Water Quality Assessment

In the late 20th century, WQIs emerged as essential tools for assessing the quality of water, particularly in rivers. In 1965, Horton was the first to ever create a water pollution control rating system based on the use of 10 variables, including DO and pH. In 1970, Brown et al., developed a WQI using 9 variables, determining their significance by calculated mean arithmetic for each variable, and then aggregating them geometrically. In 1982 Steinhart et al., built a form of an Environmental Quality Index (EQI) specifically for the Great Lakes, which included biological, physical and toxic factors. The British Columbia WQI was developed by Canada and changed in 2001 to the CCME WQI (Canadian Council of Ministers of the Environment). The WQI for Taiwan was developed by Liou et al., and the WQI developed by Said et al., used logarithmic aggregation in Florida, which utilized fewer variables compared to other WQIs. The Malaysian WQI was developed using six variables and additive aggregation. The Hanh and Almeida indices were both created with this goal in mind. Created in Vietnam (Hanh) and in Argentina (Almeida), these new indices attempt to make improvements by decreasing redundancy among the different variables and using expert opinion to determine how much weight to give to each of the nine major factors measuring water quality. Each of these indices uses a scale that ranks water quality on a continuum from poor through to excellent [5, 6].

Figure 1.1 presents the historical development of Water Quality Indices, WQIs, from the first Horton Index in 1965 to later regional indices, such as the West Java Index in 2017. It shows how the concept of water quality assessment began in the United States and then spread

To different countries and regions, including Scotland, Canada, New Zealand, Poland, Taiwan, Vietnam, Malaysia, and Argentina.

The fundamental WQIs are original indices developed for specific regions or environmental contexts. These include the Horton Index, SRDD Index, Ross Index, Environmental Quality Index, Smith Index, Dojlido Index, British Columbia WQI, Liou Index, Said Index, Malaysia Index, and Almeida Index. The first modified versions were adapted from earlier models to fit local water conditions, monitoring needs, or national standards. For example, the NSF Index was derived from the Horton Index. Later, the Dinius Index, Oregon Index, and West Java Index were derived from or influenced by the NSF Index. This shows the central role of the NSF Index in the evolution of later WQIs.

The figure also indicates two main trends in the development of WQIs. Some indices were created as independent tools for specific countries or regions, while others were produced by modifying previous models. This reflects the need to adjust WQI methods according to local environmental conditions, available parameters, and water management objectives.

Overall, the figure highlights the gradual evolution of WQIs from general models toward more localized and adapted indices. It suggests water quality assessment has become more region-specific over time, since each country or basin requires an index suited to its own physical, chemical, biological, and regulatory conditions.

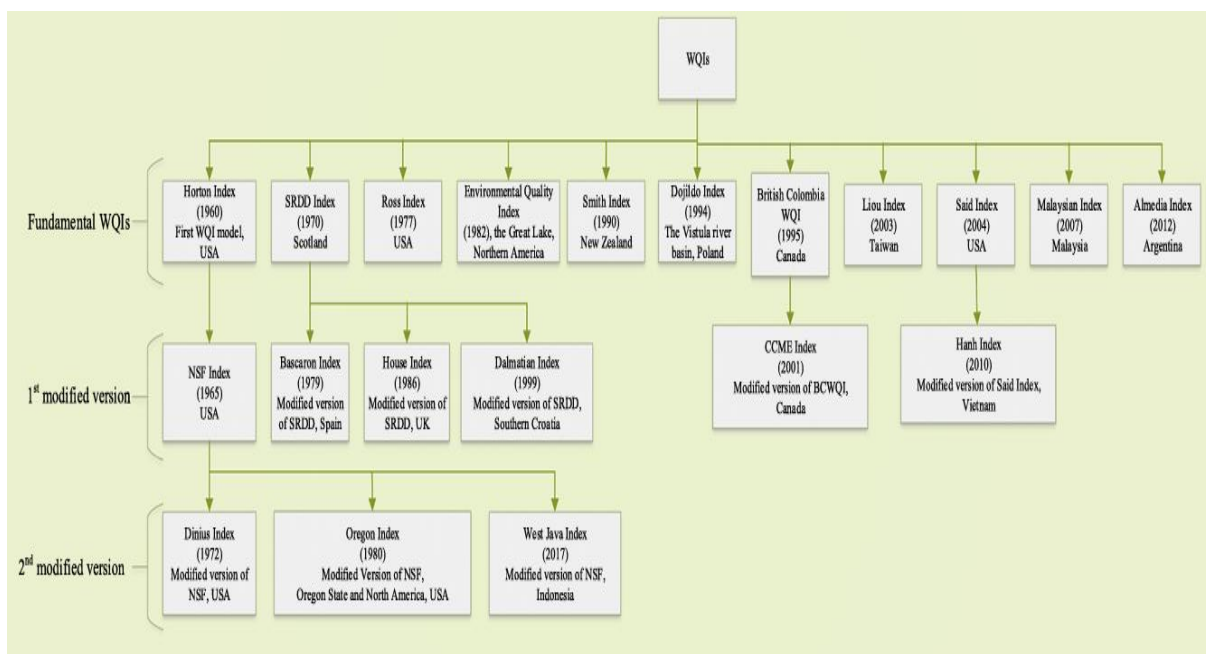


Figure 1.1. Historical development of the WQI model [11].

I.3 Different Uses of Water

Water is vital to many facets of daily life and economy. It is required for domestic tasks like cooking, drinking, cleaning, and personal hygiene. In agriculture, water helps irrigation and livestock care. Solvents in chemical processes, manufacturing, and cooling systems are examples of industrial uses. Water is also crucial to the production of energy, particularly for power plant cooling and hydroelectric power. Additionally, it promotes human well-being, ecosystem preservation, and leisure and environmental activities [7].

I.4 Definition of Water Quality Index

Water Quality Indicators (WQIs) are mathematical constructs that combine sets of data and parameters to produce a single numerical or dimensionless measure known as a composite index. The values of WQIs are typically between 0 and 100. WQIs also can provide information about various aspects of water quality (such as qualitative parameters) within established categories and rank water quality as poor, good, or excellent. WQIs provide a broad overview of the water quality, while making relevant data publicly available. They are useful in managing water resources, watersheds, and the treatment of wastewater. WQIs improve water quality reporting and assist decision making.

I.4.1 Categories of WQI

An essential first step in comprehending how water resources are assessed and their uses are appropriately directed is the classification of the water quality index. Because water monitoring has a variety of goals, different models with different mathematical philosophies and criteria have been devised. These indices can be broadly classified into four kinds, from those that rely on expert opinion to those that use statistical analysis to guarantee objectivity and lessen human bias.

I.4.1.1 Public Indices

These indices, such as the (NSFWQI), are used for overall water quality and disregard the type of water usage within the study procedure.

I.4.1.2 Specific Consumption Indices

The Oregon and British Columbia indices, for example, classify water based on its use and consumption for drinking, industry, and ecosystem preservation, among other purposes.

I.4.1.3 Designing/Planning Indices

Third, creating or organizing indices: This class serves as a tool for organizing water quality management initiatives and supporting decision-making.

I.4.1.4 Statistical Indices

These indices are based on statistical techniques and do not take into account individual opinions. Here, statistical methods are employed to assess the data. Relevance-bound assumptions of water quality observations are a crucial component of the statistical technique. [8]

I.4.2 Methodological Framework for WQI Calculation

The methodology for calculating WQI has been modified by scientists and experts.

A popular WQI approach involves the following steps (Figure 1.2):

I.4.2.1 Water Quality Measurement Parameter Selection

Parameter selection is based on management objectives and the environmental characteristics of the study area. Since numerous variables significantly impact water quality, it is recommended to use variables categorized into five main groups: oxygen content, eutrophication, health factors, physical properties, and dissolved substances.

I.4.2.2 Standardized Scale Transformation of Raw Data Parameters

Various statistical methods can be used for parameter transformation. All parameters are converted from raw data of different dimensions and units to a standardized scale. This process generates a dimensionless scale and produces sub-indices.

I.4.2.3 Providing weights to the parameters

Weights are assigned to each parameter based on its importance and impact on water quality. The determination of weights requires expert consultation. Weights are based on drinking water standards established by international and national authoritative organizations.

I.4.2.4 Combining Sub-Indices Using a Mathematical Expression

This is often done with arithmetic or geometric averages [6].

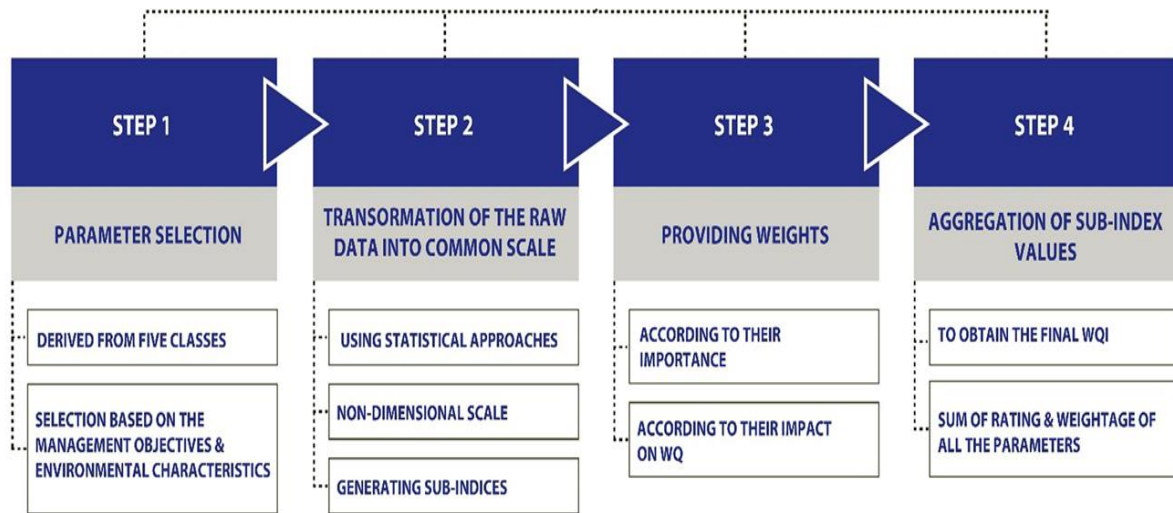


Figure 1.2. Phases of WQI development [9].

I.4.3 Different methods for WQI determination

Water quality is assessed using water quality indices. It is necessary to have a basic understanding of water quality in order to apply these indices. The National Sanitation Foundation Water Quality Index (NSFWQI), the Council of Ministers of Environment Canada Water Quality Index (CCMEWQI), the Oregon Water Quality Index (OWQI), and the Weighted Arithmetic Water Quality Index (WAWQI), which is the subject of this chapter, are some of the current water quality indices. Water quality is assessed using water quality indices. It is necessary to have a basic understanding of water quality in order to apply these indices. The NSFWQI, the CCMEWQI, the OWQI, and the WAWQI, which is the subject of this chapter, are some of the current water quality indices. The standards used to make these indices are not the same as regional standards, and they are used differently in each region. They also help show the yearly cycles, trends, and changes in water quality over time and space. These indices show the quality of the water in lakes, streams, rivers, and reservoirs in other words [10].

I.4.3.1 National Sanitation Foundation Water Quality Index (NSFWQI)

Brown developed the NSF WQI in 1965 as a modified version of the Horton model. It has been used to assess surface water quality in multiple domains [11].

The NSFWQI is a frequently used indicator that measures nine parameters: pH, total solids, fecal coliform, total phosphate, dissolved oxygen, BOD5, nitrate, turbidity, and temperature [12]. NSFWQI is calculated as follows:

$$NSFWQI = \sum_{i=1}^n Q_i W_i \quad (1.1)$$

Where,

Q_i = sub-index for i th water quality parameter;

W_i = weight associated with i th water quality parameter;

n = number of water quality parameters.

1.4.3.2 Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI)

The CCMEWQI approach simplifies difficult and technical information. The multi-variable water quality data is tested and compared to user-determined criteria. The sampling methodology needs at least four parameters to be sampled four times, but does not specify which ones to use, leaving it up to the user to choose. However, parameters may vary across stations [13]. To calculate index scores in the CCME WQI approach, use the following relation:

$$WQI = 100 - \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \quad (1.2)$$

Where, Scope (F_1) = Number of variables, whose objectives are not met.

$$F_1 = \frac{(\text{Number of failed variables})}{(\text{Total number of variables})} * 100 \quad (1.3)$$

Frequency (F_2) = Number of times by which the objectives are not met.

$$F_2 = \frac{(\text{Number of failed tests})}{(\text{Total number of tests})} * 100 \quad (1.4)$$

- The word "excursion" refers to when the number of an individual parameter exceeds the minimum (when the target is less than).

a) When the test value must not exceed the objective:

$$excursion_i = \left(\frac{\text{Failed test value}_i}{\text{Objective}_i} \right) - 1 \quad (1.5)$$

b) When the test value must not fall below the objective:

$$excursion_i = \left(\frac{\text{Objective}_i}{\text{Failed test value}_i} \right) - 1 \quad (1.6)$$

$$\text{Normalized sum of excursions } nse = \frac{\sum_{i=1}^n excursion_i}{\text{number of tests}} \quad (1.7)$$

Amplitude (F3) = Amount by which the objectives are not met [14].

$$F3 = \frac{nse}{0.01nse+0.01} \quad (1.8)$$

I.4.3.3 Oregon Water Quality Index (OWQI)

Describes the quality of surface water bodies. This descriptive technique for water quality uses eight physical, chemical, and biological parameters: temperature, dissolved oxygen (DO), biochemical oxygen demand (BOD), pH, ammonia+nitrate nitrogen, total phosphorus, total dissolved solids (TDS), and coliforms. The recommended index is based on the NSF-WQI model. The distinction resides in the calculus approach and the weight of parameters, which are not considered in this scenario. This method's mathematical representation is based on the notion of arithmetic average [15] and is expressed using a formula:

$$OWQI = \frac{n}{\sqrt{\sum_{i=1}^n SI^2}} \quad (1.9)$$

Where, n is the number of parameters, SI_i is the value of parameter i .

I.4.3.4 Weighted Arithmetic Water Quality Index Method (WAWQI)

The weighted arithmetic (WA) approach is often used to examine WQI. The WA technique classifies WQ based on purity using widely measurable characteristics. This approach has been frequently used. The concept of this strategy is based on the weight of each parameter. The weight ranges from less than 0.000667 to more than 330. The weights vary significantly. The relative weights are calculated using values that are inversely proportional to the parameter standards [16]. WQI is calculated using the following formulae:

$$WAWQI = \frac{\sum_{i=1}^{i=n} Q_i W_i}{\sum_{i=1}^{i=n} W_i} \quad (1.10)$$

In Table 1.1 represent an overview of Water Quality Rating as per different Water Quality Index methods; we will see the details of calculating this indicator later.

Table 1.1. Water Quality Rating as per different Water Quality Index methods [17].

National Sanitation Foundation Water Quality Index (NSFWQI)	
WQI Value	Rating of Water Quality
91-100	Excellent water quality
71-90	Good water quality
51-70	Medium water quality
26-50	Bad water quality
0-25	Very bad water quality
Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI)	
95-100	Excellent water quality
80-94	Good water quality
60-79	Fair water quality
45-59	Marginal water quality
0-44	Poor water quality
Oregon Water Quality Index (OWQI)	
90-100	Excellent water quality
85-89	Good water quality
80-84	Fair water quality
60-79	Poor water quality
0-59	Very poor water quality
Weighted Arithmetic Water Quality Index Method(WAWQI)	
Above 100	Unsuitable for drinking purpose
76-100	Very Poor water quality
51-75	Poor water quality
26-50	Good water quality
0-25	Excellent water quality

I.5 Selection of Water Quality Parameters

I.5.1 Water quality standards

Evaluating the quality of water in important resources like dams and groundwater is a very important step in making sure they are safe for different uses. According to Algerian national standards and World Health Organization (WHO) recommendations, water that is consumed must not contain pathogens, toxic substances, or high levels of minerals or organic matter that could make it hard to use in farming, industry, or everyday life. It should also keep its physical properties, such as being clear and not having any bad smells or tastes. The WQI is a strong numerical tool that makes it easier to evaluate these complicated variables. This index combines different physical and chemical factors into one score, which makes it easy to see how the water is doing (from "Excellent" to "Unsuitable"). [18]

Table 1.2. Standard maximum values of certain parameters to be used in this study [19].

Parameters	Si (standard maximum value, Algeria)
pH	9
Conductivity (à 25 °C ms/cm)	2800
Temperature (°C)	25
Turbidity (NTU)	5
Total Dissolved Solids (mg/L)	500
Dissolved Oxygen (mg/L)	5
Chlorides Cl^- (mg/L)	250
Sulphates SO_4^{2-} (mg/L)	250
Nitrates NO_3^- (mg/L)	50
Phosphates PO_4^{3-} (mg/L)	0.5

I.5.2 Water quality measurement parameters

I.5.2.1 Physicochemical parameters

I.5.2.1.1 pH Level

The acidity or alkalinity of water is indicated by the pH level. example it is commonly accepted that drinking water should have a pH between 6.5 and 8.5. Extreme pH water can contain dangerous pollutants and erode pipes [20].

I.5.2.1.2 Temperature

The temperature evolution is determined by the method of use of the water before to its discharge, as well as the path of the water prior to the exhaust or purification station. In general, water should be evacuated into the environment at temperatures below 30°C; hotter water constitutes pollution [21].

I.5.2.1.3 Total Dissolved Solids (TDS)

TDS determines the total concentration of dissolved components in water, which includes minerals, salts, and metals. Low TDS levels (less than 500 mg/L) are preferable for drinking water, however higher levels may impair flavor and usability [20].

I.5.2.1.4 Electrical Conductivity (EC)

EC assesses the capacity of water to conduct electricity, which is proportional to the concentration of dissolved ions (salts). It serves as an indirect indication of TDS levels. High EC values indicate high mineral concentration [20].

I.5.2.1.5 Dissolved oxygen

The Water Quality Index (WQI) uses dissolved oxygen (DO) as a key measure of water quality because it shows how healthy the water is for living things.

If the DO levels are high, the water is good; if they are low, the water is getting worse because of organic pollution. If the DO level goes down, the WQI value goes down too, which means the water is of lower quality. On the other hand, if the DO level goes up, the WQI value goes up, which means the water is good for aquatic life [21].

I.5.2.2 Pollution parameters

I.5.2.2.1 Sulfate

Sulfates come from the natural breakdown of some minerals, like gypsum. They can also come from human activities, like industrial discharges (mining, metal processing), wastewater, burning trash, or leachate from landfills. The most sulfates that can be in water is 250 mg/L. Concentrations above this limit are not considered dangerous, but they may cause stomach problems, especially diarrhea, especially in kids. Sulfates are chemical elements that are often checked in reservoir and well waters because they can be harmful at high levels [21].

I.5.2.2.2 Nitrate Concentration

Agricultural runoff and sewage are common sources of nitrates in groundwater. High amounts (more than 50 mg/L) can be dangerous, especially to newborns, and are an indication of contamination [20].

I.5.2.2.3 Chloride Concentration

Chloride levels indicate salty water infiltration or pollution due to sewage and industrial activity. Elevated chloride levels can render water unpleasant and unsuitable for irrigation [20].

I.6 Water Quality Index (WQI)

The Water Quality Index (WQI) is a standard way to check the quality of water and see if it is safe to drink. The most common way to check the quality of water is to use WQI. The weighted arithmetic index was calculated using the weighted average (WA) method. To figure out the WQI, you need to give different physical and chemical parameters weights (w_i) based on how they affect water quality. There are four main steps to WQI: choosing parameters, giving those weights (w_i) and calculating weights (W_i), calculating sub-indices (S_i), and finally adding them all together to get the final WQI. The weights given range from 1 to 5, with higher weights meaning a bigger effect on water quality. We figured out the WQI by doing the following:

Step 1: The quality rating scale of Sub-indices (q_i) for all variables, used the following equation to figure out:

$$q_i = \frac{C_i}{S_i} \times 100 \quad (1.11)$$

Where;

- The sub-index q_i is the quality rating for the i th parameter.
- C_i is the measured concentration value (mg/l) of each parameter.
- S_i is the standard maximum value of each parameter.

For pH and DO parameters specifically, where V_i = Ideal value which is considered as 7.0 for pH and 14.6 mg/l for DO, the q_i was calculated by using given equation:

$$q_i = \frac{C_i - V_i}{S_i - V_i} \times 100 \quad (1.12)$$

Step 2: The following equation was used to figure out the relative weight for each parameter W_i :

$$W_i = \frac{k}{S_i} \quad (1.13)$$

Where k is the constant of proportionality, and can also find it using the following equation:

$$k = \frac{1}{\sum_{i=1}^n \frac{1}{S_i}} \quad (1.14)$$

The relative weight W_i can calculate also by another equation:

$$W_i = \frac{w_i}{\sum_{i=1}^n w_i} \quad (1.15)$$

Where; W_i is the relative weight w_i is the proposed and selected weight of each parameter and n is the total number of parameters.

Step 3: To calculate the water quality index (WQI), first determined the sub-index of each parameter and the relative weight, with the final calculation using the following equation:

$$SI_i = W_i \times q_i \quad (1.16)$$

General formulae:

$$WQI = \sum_{i=1}^n SI_i \quad (1.17)$$

Where SI_i the sub-index of the i th parameter, and n is the total number of variables considered.

Table 1.3. Classification of water quality based on the WQI scale and corresponding water status [22]

WQI Range	Water quality Class	Description
0 - 25	Excellent	Drinking water, Irrigation and industry
26-50	Good	Drinking water, Irrigation and industry
51-75	Poor	Irrigation and industry
76-100	Very poor	Irrigation
>100	Unsuitable for drinking	Proper treatment

I.7 Justification for use of WQI

Are used to review water quality monitoring data and provide considerable interpretation of results, particularly when pollutant concentrations are below water quality criteria. WQI may be used to monitor water quality without the need for frequent sample. WQIs enable specialists to contextualize monitoring data, analyze the success of regulatory initiatives, and efficiently communicate water quality information to stakeholders. Indices were used for many monitoring objectives, including water quality evaluation, treatment, usage, public information, research and development, and environmental planning [23, 24].

I.8 Conclusion

This chapter provided a comprehensive overview of water quality monitoring. The Water Quality Index is a useful and simple technique for reporting water quality data. It has proven effective in evaluating temporal and spatial changes in water quality in different regions worldwide. Water quality indices convert large and complex water quality datasets into a single value in a simple and reproducible way. Their wide and successful application over the past 60 years demonstrates their importance in water quality assessment. They help summarize large amounts of complex scientific data and clearly describe water quality status to the public. Even non-technical stakeholders understand the simple dimensionless index score.

The next chapter focuses on the theoretical aspects of Machine Learning implementation. It examines how Machine Learning, when applied to multi-sensor data, provides a reliable approach for constructing predictive models to improve water quality evaluation through pattern identification.

Chapter II

Machines Learning

Methods

II.1 Introduction

Machine learning has greatly transformed water quality assessment by replacing manual calculation methods with automated, adaptable computational techniques. Unlike traditional approaches that depend on fixed mathematical formulas, machine learning algorithms—such as Random Forest (RF) and Support Vector Regression (SVR), can effectively model complex relationships among water quality parameters like pH and dissolved oxygen. These algorithms facilitate accurate prediction of the WQI using historical data, enhancing real-time monitoring capabilities and enabling forecasting of future water pollution trends.

II.2 Study Area Description

II.2.1 Geographic Location

The Tilesdit Dam is located in northern Algeria, approximately 18 km east of Bouira city and 4 km from National Road No. 5. Geographically, the dam is situated on Oued Eddous at the outlet of a 843 km² watershed [25]. As shown in Figure II.1, the latitude and longitude of the study area are 35° 13' 22" N and 4° 14' 23" E, respectively.



Figure 2.1. Geographic location of Tilesdit dam in Bouira, Algeria.

II.2.2 Climate Characteristics

The climate of the study area significantly influences the hydrological cycle and the water quality in the Tilesdit Dam. Based on regional geographic data, the climate is classified as **Semi-Arid Mediterranean**, characterized by:

- **Seasonal Variation:** The area experiences two main seasons—a hot, dry summer and a cold, moist winter—causing notable fluctuations in the water’s physical and chemical properties.

- **Precipitation:** Water inflow depends on annual rainfall averaging **500–600 mm** in this part of northern Algeria.

- **Temperature Influence:** High summer temperatures increase evaporation rates, potentially raising the concentrations of minerals and pollutants in the reservoir [25].

II.2.3 Water Source Description

The reservoir is designed to store water with a total capacity of 167 million m³ and a surface area of 843 hectares. The average annual water inflow is estimated at 65 million m³ the dam serves a dual purpose: providing 23 million m³ for potable and industrial water supply, and 50 million m³ for agricultural irrigation. The water flows from Oued Eddous into Oued Sahal, eventually joining the Soummam basin [25].

II.3 Data Sources and Sensors Information

The reliability of any machine learning model depends on the quality and consistency of the input data. For this study, the dataset was acquired through rigorous daily monitoring at the Tilesdit Dam. The technical services department at the dam utilized high-precision instrumentation and automated sensors to record the physical and chemical properties of the water.

The raw data consists of a comprehensive set of various parameters, which are essential for understanding the ecological and chemical status of the reservoir. These parameters include:

- **Physical Parameters:** Temperature, Turbidity, and Electrical Conductivity (EC).
- **Chemical Parameters:** pH, Dissolved Oxygen (DO), Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), and Suspended Solids (SS).

The use of continuous sensing technology ensures that the dataset captures the dynamic changes in water quality caused by seasonal transitions and human activities. These 800 observations were later organized into a digital format, forming the basis for training the SVR, Random Forest, and ANN models discussed later in this chapter.

II.4 Data Preprocessing Strategy

II.4.1 Data Cleaning

Cleaning up the Data before we can train our machine learning models, we need to make sure our raw data is accurate and reliable. To do this, we follow a three-step cleaning process. First, we deal with missing values. Sometimes, our sensors don't work properly or there are errors when we're sending the data, which means we're left with gaps in our data. To fix this, we use a method called Linear Interpolation to fill in the missing values. This helps us keep the timeline of our data intact, which is really important when we're trying to forecast what will happen in a certain area over time. Next, we look for outliers. These are measurements that don't seem right, like a pH value that's way outside the normal range or a sudden spike in turbidity that's probably just a mistake. We use something called the Z-score method to find these outliers, and if they don't make sense in the real world, we get rid of them so they don't mess up our models. Finally, we make sure everything is consistent. Since we're getting data from lots of different places, we need to make sure all the units of measurement are the same and that our clocks are all synchronized. This way, we can be sure that our data is reliable and will work well with our models. By following these steps, we can be confident that our data is clean and ready to use. This is a crucial part of our process, as it helps us build models that are accurate and useful [26].

II.4.2 Data Normalization and Standardization

The Data Normalization and Standardization is an important step when we are working with machine learning (Figure 2.1). This is especially true when we are dealing with water quality indices that're not the same. The Data Normalization and Standardization helps to make sure that all the parameters are treated equally by the model.

The dataset has features with different values. For example the pH values are between 0 and 14. Other parameters can be in the thousands. If we do not scale these features the algorithms like Support Vector Regression or Neural Networks may not work properly. They may give importance to the features with bigger ranges. This can lead to poor results and bad predictions.

To avoid this problem we use two methods.

* **Normalization:** this method changes the data so that it is within a range usually between 0 and 1. This method works well when the data does not follow a normal distribution.

* **Standardization:** this method changes the data so that it is around a mean of 0 and a standard deviation of 1. This method is very good at handling outliers, which's very important when we are dealing with data, from sensors that monitor dams.

Using these methods makes the model more stable. Reduces the time it takes to train. It also makes sure that the features are balanced across the 37 x 11 matrix. The Data Normalization and Standardization is a step that helps the model to work properly. [27]

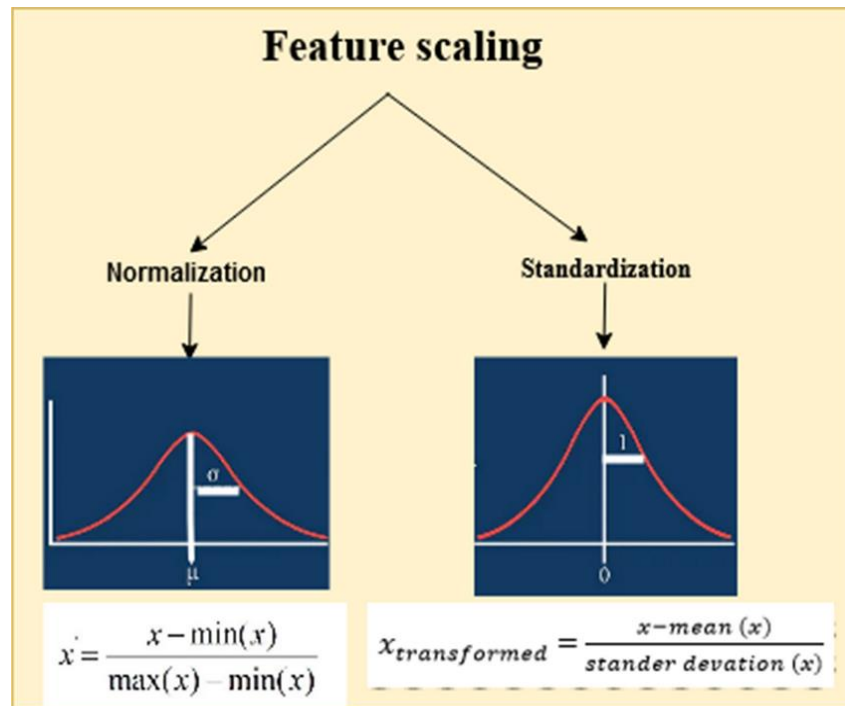


Figure 2.2. Mathematical formulation and visual distribution of Normalization versus Standardization methods.

II.4.3 Correlation Analysis of Water Quality Parameters

Before feeding the **800 samples** into our machine learning models, it is essential to understand the statistical relationships between the **8 physicochemical parameters**. We conducted a Correlation Analysis using the **Pearson Correlation Coefficient** to identify how different variables (such as Temperature, Turbidity, and Conductivity) influence the target pH or WQI [46].

This step is crucial for three reasons:

1. **Redundancy Reduction:** It helps identify if two sensors are giving nearly identical information, allowing us to simplify the model.

2. **Feature Importance:** It reveals which variables have the strongest impact on water quality.

3. **Data Insight:** For the **800 observations**, a correlation "Heatmap" provides a visual summary of the reservoir's chemical behavior over the study period.

By analyzing the correlation matrix, we ensure that the RF and SVR models are trained on the most significant features, leading to higher accuracy and lower error rates [28].

II.5 Machine Learning Framework

II.5.1 Support Vector Regression (SVR)

1.1. Support Vector Regression

Support Vector Regression (SVR) is an extension of Support Vector Machines (SVM) developed by Vapnik, commonly used for regression and classification purposes. The fundamental purpose of SVR is to identify an optimal hyperplane by looking for a function $f(x)$ that deviates from the target values y by no more than ε , while maximizing the margin, employing an ε -insensitive loss function that ignores mistakes less than ε . Figure 2.2 illustrates how the SVR model fits the data, and a linear relation represents the nonlinear function $f(x)$ of the hyperplane:

$$f(x) = \langle w, x \rangle + b = y ; w \in R^n, b \in R \quad (2.1)$$

Where w and b are the coefficients that have to be estimated from data.

Any function that satisfies Mercer's theorem can be used as a kernel function to compute a dot product in feature space. It employs a high-dimensional transformation through kernel nonlinear functions $\phi(x)$.

$$\begin{cases} \max L(\alpha_i, \alpha_i^*) = -\frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \phi(x_i), \phi(x_j) \rangle - \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*) \\ \text{with : } \begin{cases} \sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C \end{cases} \end{cases} \quad (2.2)$$

The Gaussian kernel, often known as the Radial Basis Function (RBF), is the most widely utilized kernel:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (2.3)$$

The three main SVR parameters are C , ε , and kernel parameters. C regulates regularization, ε establishes the insensitivity margin, and the Lagrange multipliers α_i and α_i^* modify the predictions in the direction of the goal value. The unrealizable constraints of the convex optimization problem, also known as the soft margin, can be managed by introducing spring variables (or relaxation variables) using the linear ε -insensitive loss function $f(x)$ (Figure 2.2-a). Figures 2.2-b and 22-c depict the linear ε -insensitive loss function, which is expressed as follows:

$$l(\xi)_\varepsilon = \begin{cases} 0 & \text{si } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \quad (2.4)$$

such as $\xi = y - f(x)$

The final decision function is expressed as:

$$f(x, \alpha, \alpha^*) = \sum_{i=1}^{nsv} (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (2.5)$$

where nsv is the number of support vectors contribute to the prediction.

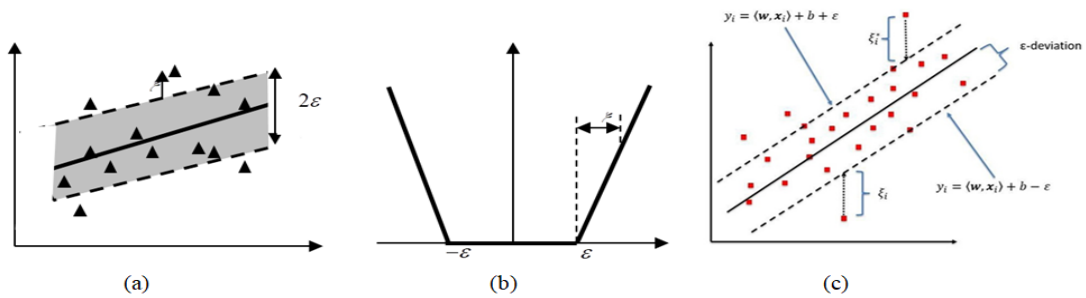


Figure 2.3. SVR Predictive Function Components: (a) Soft Margin, (b) ε -Insensitive Loss function, (c) Optimal hyperplane and Support Vector.

Support Vector Regression is a useful way to predict things like water quality parameters. This is because it can deal with things that're not straight forward. For example it can handle things that're not linear.

According to some people who did a study Support Vector Regression is very good at predicting water quality. This is because it uses something called Kernel. These Kernel Functions can take the information we have. Turn it into something that is more useful. They can make it easier to find the way to predict things.

Some people have used Support Vector Regression to study rivers. They have tried Kernel Functions like the RBF and the Pearson VII Universal Kernel. They found out that using these Kernel Functions with Support Vector Regression can give good results. It can even be better than using methods.

In our study we used Support Vector Regression because it can help us make sure our results are good and will work in situations. This is really important when we are dealing with a lot of information, from sensors [29].

II.5.2 Random Forest (RF)

- The Random Forest method is a way of learning that works by making a lot of decision trees when it is being trained. The Random Forest is strong because it can put together what it learns from individual trees to get a better and more stable result.
- The Random Forest method was first talked about by Breiman [30]. In the case of studying water Al-Nahid [31] and others showed that the Random Forest is very good at predicting the quality of water because it does not get confused by much data and it can handle complicated relationships in big datasets from the environment. The Random Forest uses a technique called Bootstrap Aggregating where each tree is trained on a set of data. This helps to make sure the trees are different and it reduces the differences in the model.
- For the water quality dataset the Random Forest model is used to understand the relationships, between the 8 physical and chemical features and the pH values we are trying to predict. The Random Forest method gives us a way to predict what will happen and it works well with the SVR approach.

The figure 2.3 explains a random forest regression workflow:

Source data first becomes measured data, and then the dataset is split into two parts: 80% for training and 20% for testing or validation.

- In Step 1, the random forest is built. The training data is divided into several subsets. Each subset trains one decision tree. These trees work independently, so Decision tree 1, Decision tree i, and Decision tree n learn different patterns from the data.
- In Step 2, each decision tree gives its own prediction. Since this is regression, the final prediction comes from combining the outputs of all trees, usually by averaging their predicted values. This reduces the error compared with relying on one decision tree.

After prediction, the model performance is evaluated using the test data. The figure uses MSE, Mean Squared Error, as the performance metric. A lower MSE means the predicted values are closer to the measured values.

- In Step 3, feature contribution ranking is calculated. This step identifies which input variables had the strongest influence on the model prediction. In random forest regression, feature importance is often estimated by measuring how much each feature reduces prediction error across the trees.

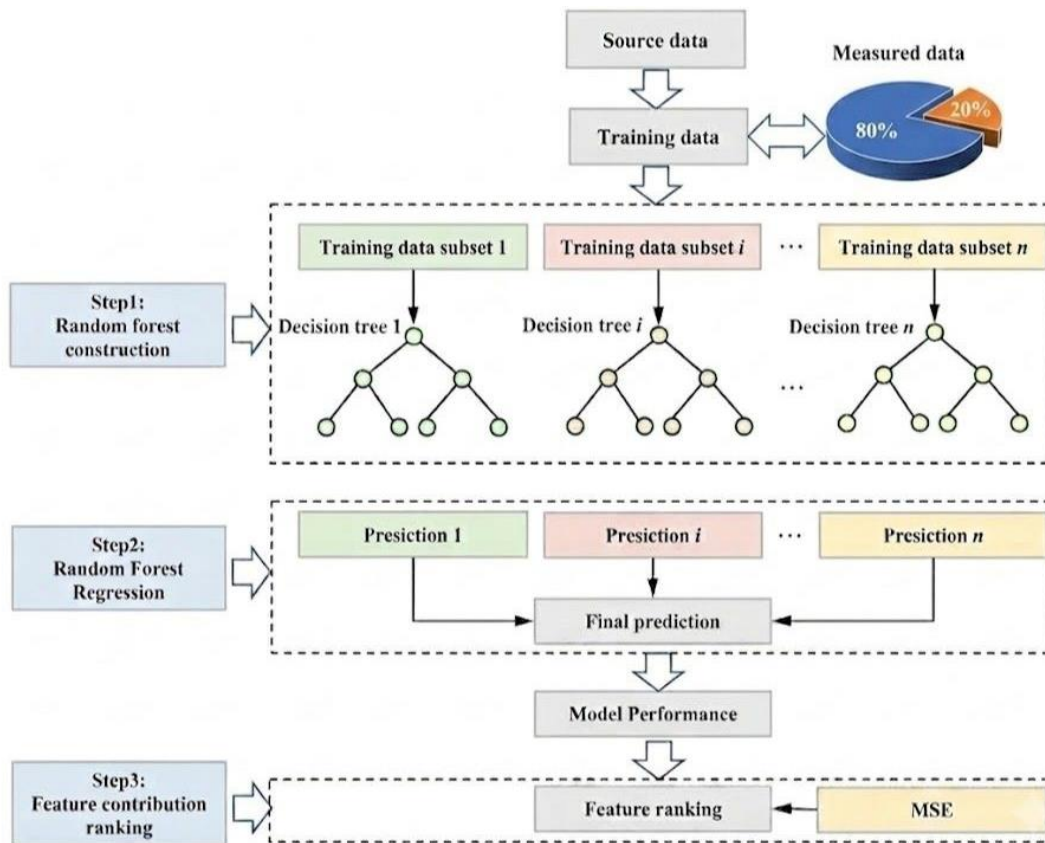


Figure 2.4. Workflow of Random Forest Regression and Feature Importance Ranking.

II.6 Feature Importance and selection Strategy

We use a feature selection strategy to pick the input data for predicting WQI values. This approach, used by Latifoğlu and Doğan helps simplify the data.

Hydrological data often has variables and changes over time. So we need to choose the important features. Research on river basins shows that using variables with a RFR-based selection method gives accurate predictions.

By using this strategy we focus on the relevant features related to space and time. This helps in ways:

- It makes our model less complicated. Prevents it from fitting too closely to the training data.
- It finds the significant time delays between water quality parameters and the pH value we are trying to predict.
- It supports reliable decisions for managing water at the water quality monitoring system.

We apply the feature selection strategy to achieve these benefits and improve prediction. The strategy helps us get results, similar, to previous studies that achieved best results [32].

II.6.1 Importance for Instrumentation and Machine Learning

The use of Machine Learning in instrumentation systems is a step forward for environmental monitoring. At the water quality monitoring we can see why this combination is important in ways:

- **Sensor Data Fusion:** Instrumentation systems generate complex data, and machine learning identifies patterns that traditional methods might overlook by processing sensor inputs via SVR and Random Forest.
- **Non-linear Calibration:** While many sensors assume linear relationships, water parameters like pH and turbidity are non-linear; machine learning models these complexities without manual sensor adjustment.
- **Virtual Sensing (Soft Sensors):** Machine learning enables the creation of "soft sensors" to predict difficult-to-measure chemical levels using basic parameters like temperature, reducing the need for expensive hardware.
- **Noise and Fault Tolerance:** These models enhance system reliability by detecting sensor faults or transmission errors, ensuring continuous monitoring of water quality [32].

II.7 Model Training and Validation

II.7.1 Training Process

The training process is really important because this is where the machine learning model figures out how to connect the dots between the input variables and the Water Quality Index (WQI).

According to Ehteram (2024) the training process has a key steps that make it work well.

- **Data Preprocessing:** First we take the data from sensors like pH, Turbidity and DO and clean it up so the learning algorithm can use it. This step is crucial because the model needs data to find meaningful patterns in the water quality parameters.
- **Layer-wise Feature Extraction:** In the architectures the training starts with convolutional layers that look for special features and patterns in the data. Then we use activation functions like the Rectified Linear Unit (ReLU) to add some complexity to the model.
- **Recursive. Temporal Dependencies:** For the hybrid models the training process uses layers like CRNN that help the network learn about short-term and long-term patterns in the water resource data.
- **Optimization via Backpropagation:** We adjust the models parameters, like the weights and biases to minimize the loss function. We use the Adam optimizer to do this, which helps the model learn at the pace.
- **Stopping Criteria:** The training keeps going until it meets a condition like a low enough error rate or a set number of iterations which means the Water Quality Index model is ready to make good predictions [33].

II.7.2 Cross-Validation Method

To ensure the reliability and robustness of predictive models, a systematic validation approach is essential. As noted by Ehteram et al. (2024), K-fold cross-validation (KFCV) is a highly effective technique for assessing model performance and determining suitable training and testing data sizes.

The key steps in implementing KFCV are:

- **Dataset Partitioning:** The dataset is split into K equal-sized subsets or "folds."
- **Iterative Training and Testing:** In each iteration, the model is trained on (K-1) folds and validated on the remaining fold.
- **Performance Averaging:** Performance metrics from all K iterations are averaged to provide an unbiased overall estimate.
- **Risk Reduction:** By using multiple validation sets, KFCV reduces overfitting and under fitting risks, making the model less sensitive to data partitioning.

This rigorous validation process helps ensure that the predictive system for the water quality monitoring accurately forecasts the Water Quality Index (WQI) for new, unseen environmental data [34].

II.7.3 Testing and Evaluation Metrics

To quantify the predictive accuracy and efficiency of the developed models, several statistical evaluation metrics are employed. According to Ehteram et al. (2024), these metrics provide a comprehensive assessment of the model's ability to forecast the Water Quality Index (WQI) by comparing predicted values with observed data:

- **Coefficient of Determination (R^2):** This metric indicates the proportion of variance in the dependent variable that is predictable from the independent variables. A value closer to 1.0 signifies a near-perfect correlation between the sensor inputs and the predicted WQI.
- **Mean Absolute Error (MAE):** Without taking into account the direction of the errors, it calculates the average size of the errors in a set of forecasts. Higher precision in the monitoring system is indicated by lower MAE values.
- **Root Mean Square Error (RMSE):** This is a standard way to measure the error of a model in predicting quantitative data. It is particularly sensitive to outliers, ensuring that any significant deviations in water quality are accurately captured [32].

Table 2.1. Mathematical Formulation of Performance Evaluation Metrics [35].

etrics	athematical Formula	Description
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	Measures the proportion of variance in the dependent variable explained by the model
Mean Absolute Error (MAE)	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	Measures the average magnitude of the errors in a set of predictions, without considering their direction.
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	Measures the standard deviation of the prediction errors, emphasizing large errors

II.8 Conclusion

This chapter has established the theoretical and methodological foundation for the intelligent prediction of the Water Quality Index (WQI). By exploring machine learning architectures such as Support Vector Regression (SVR) and Random Forest, we identified key algorithms capable of managing the non-linear complexities inherent in environmental data.

The adoption of the K-fold cross-validation method alongside rigorous performance metrics—including (R^2), MAE and RMSE, provides a robust framework for model training and validation. These approaches ensure that the predictive system is both accurate and generalizable to new data.

This foundational framework sets the stage for the next chapter, where these models will be implemented and tested using the Tilesdit Dam dataset to identify the most effective solution for real-time water quality monitoring.

Chapter III

Results & Discussion

III.1 Introduction

Water is a fundamental element for human life and health. It transports essential nutrients and helps regulate physiological functions. In its natural state, water is rarely pure, since it often contains physical, chemical, and biological substances that require strict monitoring and treatment to meet safety standards. Water quality assessment has a direct impact on public health. It depends on several qualitative and quantitative variables, especially when water is intended for drinking or irrigation. For this reason, strong mathematical methods are needed to combine the physicochemical properties of water into a single indicator that represents its overall water quality. The WQI, provides this type of assessment. It expresses water quality as a single dimensionless value calculated from a set of measured water parameters.

The main objective of this study is to examine the different parameters, complexities, and computational requirements involved in WQI analysis. Estimating the Water Quality Index without direct analytical calculation is a pattern-recognition problem, since the model must identify relationships between several physicochemical variables and the final water quality class or value. In this context, the study presents the main steps required to calculate the WQI and develop a machine learning model capable of predicting it.

This chapter focuses on the practical application of the methodology established in the previous sections. Using real data collected from the Tilesdit Dam, the study aims to implement an intelligent water quality monitoring system. The next sections describe the study area, the physical and chemical parameters measured, and the application of ML models to predict the WQI, in accordance with World Health Organization (WHO) guidelines and Algerian national standards.

III.2 Data Collection and Sampling:

III.2.1 Sampling Strategy

Our study is based on a large dataset that was collected for the purpose of modeling water quality. The dataset comprises 800 surface water samples, subjected to measurement and physicochemical analysis to guarantee an accurate depiction of the water resource status. The database contains six basic variables or physicochemical indicators that are used as inputs for the predictive models, as shown in Figure 3.2 and Table 3.1 provides a summary of descriptive statistics for used water quality variables:

- **pH:** To measure the acidity or alkalinity of the aqueous environment.

- **Electrical Conductivity (EC):** Reflecting the concentration of dissolved salts.
- **Temperature:** Which directly affects chemical and biological reactions in the water.
- **Turbidity:** To evaluate water clarity and the presence of suspended solids.
- **Total Dissolved Solids (TDS):** A measure of the total dissolved organic and inorganic substances.
- **Dissolved Oxygen (DO):** A vital indicator for the health of aquatic ecosystems.

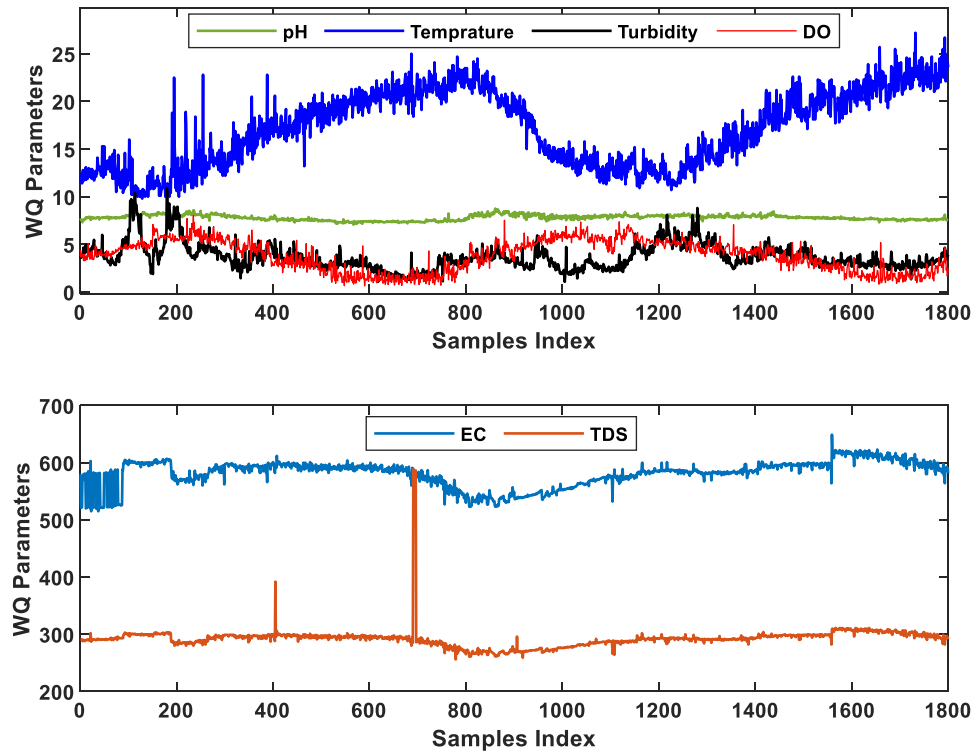


Figure 3.1. Evolution of descriptive parameters from 2016 to 2018.

Table 3.1. Descriptive statistics for used water quality variables.

Parameters	Statistical measure			
	Minimum	Maximum	Mean	Standard Deviation
Temp (°C)	9,80	27,20	12,16	1,28
pH	7,07	9,80	7,87	0,18
EC (μS/cm)	515,00	649,00	568,59	33,47
Turb (NTU)	0,76	23,81	4,93	1,81
TDS(mg/L)	256,00	589,00	294,19	4,48
DO (mg/L)	0,64	8,05	4,31	0,53

This dataset is the basis for making the suggested decision support tool. It was first used to figure out the WQI using the standard analytical method. Then, it was used to train and test ML algorithms, to make sure that a highly accurate predictive model was created. Measured Parameters:

To find out how good the water is, you need to measure a set of core physicochemical parameters and compare them to the WHO international limits. These indicators are very important for describing the chemical and physical properties of the water environment. Each one shows a different part of pollution or ecological balance. For the sake of statistical calculations, each parameter is given a weight (w_i) based on how much it affects water quality. Where the relative weight (w_i) which is the weight of a parameter shows how much that parameter affects public health. The weights are given numbers from 1, to 5. The parameters that can hurt people's health more get relative weights and this is decided using the rules of the World Health Organization. For example, Dissolved Oxygen is a very important indicator, so it gets a higher weight. This table shows a summary of these parameters, what they do, and the reference values that were used:

Table 3.2. Standard values and relative weights of water quality parameters.

Parameter	Description	Role in Water Quality	WHO Standard
pH	Measure of water acidity or alkalinity (0–14 scale)	Controls chemical reactions and affects aquatic life; indicates possible pollution	6.5 – 8.5
TDS(mg/L)	Total concentration of dissolved substances (salts, minerals, organic matter)	Indicates mineralization and pollution level; affects taste and usability	≤ 500 (desirable)
Temperature (°C)	Degree of heat in water	Influences chemical reactions and dissolved oxygen levels	$\leq 25^{\circ}\text{C}$ (recommended)
DO(mg/L)	Amount of dissolved oxygen in water	Essential for aquatic life; low values indicate pollution	≥ 5 mg/L
Conductivity ($\mu\text{S/cm}$)	Ability of water to conduct electricity due to ions	Reflects salinity and dissolved ion concentration	≤ 1500 $\mu\text{S/cm}$
Turbidity (NTU)	Measure of water clarity (suspended particles)	Indicates suspended particles and contamination	≤ 5 NTU

III.2.2 Water Quality Index Calculation

The Water Quality Index (WQI) is an effective instrument for communicating water quality. It provides a consistent unit of measurement, making it an important feature in surface water management. Quality of water indices (WQIs) are used to measure water's appropriateness for various purposes. These indicators indicate the level of water quality in lakes, streams, rivers, and reservoirs. Most WQIs compare determinants to regulatory criteria [36]. There are many methods for water quality index quantification. In this study methods was selected: the WA-WQI.

The WQI was used to aggregate various variables and their dimensions into a single value from water quality dataset of Tilesdit Dam. The calculation and formulation of the WQI

involved the three steps mentioned above, according to the equations 1.11 to 1.15 and Table 1.3. The dataset on water quality provided by the Tilesidt station was used in this analysis.

From Tables 3.3 and 3.4, the weight (W_i) of each of the six parameters was calculated using Equation 3 and the standard set (S_i) for each parameter, according to the guidelines. After calculating the WQI, the values ranged from 57,69 to 125.99. Most of the WQI values fell between "Good" and "Poor," with a few instances categorized as "Unsuitable for Drinking". An example of the calculation of the WQI for these physicochemical parameters and tis classes is shown in table 3.5. The obtained WA-WQI values are shown in Figure 3.3.

Table 3.3. Relative weights and standard values of water quality variables used in WQI calculation.

Parameters	WHO Standard	Algerian Standard	S_i Standard maximum value, Algeria	Relative Weight (W_i)
Temp (°C)	20-32	25	25	0.07227153
pH	7 - 8	6.5 - 9	9	0.20075426
EC (µS/cm)	1500	≤ 2800	2800	0.00064528
Turb (NTU)	5	≤ 5	5	0.36135767
TDS (mg/L)	1000	≤ 500	500	0.00361358
DO (mg/L)	5	≤ 5	5	0.36135767

The relative weight (W_i) of each physicochemical parameter and the proportionality constant K are first calculated using the maximum values from the Algerian standard for surface water of the studied physicochemical parameters. An example of the calculation of the Water Quality Index (WQI) for these physicochemical parameters is shown in Table 3.4.

Table 3.4. Weight of physicochemical parameters and Algerian standard for surface water quality (Algerian standard for water quality, 2016 and 2018).

	pH	TDS (mg/L)	Temperature (°C)	DO (mg/L)	Conductivity (µS/cm)	Turbidity (NTU)
S_i	9	500	25	5	2800	5
$1/S_i$	0.1111	0.002	0.04	0.2	0.00035	0.2
W_i	0.20075	0.00361	0.07227	0.36135	0.00064	0.36135
$\sum \frac{1}{S_i}$	0.55346825					
$K = \frac{1}{\sum \frac{1}{S_i}}$	1.80678836					

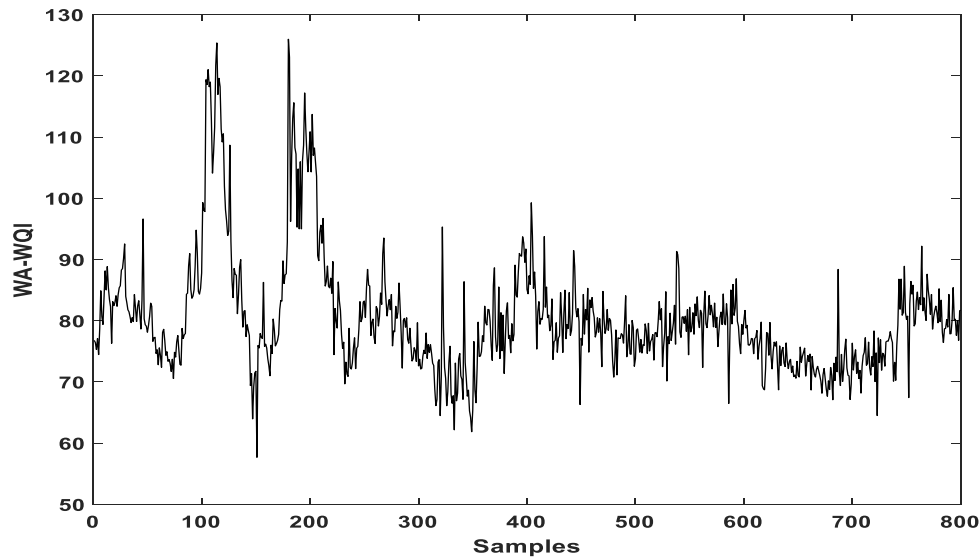


Figure 3.2. Variation in Sample Features of Six Selected WQ Parameters.

Table 3.5. An example of the calculation of the Water Quality Index (WQI)

pH	Conductivity ($\mu\text{S}/\text{cm}$)	Temperature ($^{\circ}\text{C}$)	Turbidity (NTU)	TDS (mg/l)	DO (mg/l)	WQI	WQ Class
7.74	580	11.5	3.80	291	4.41	76.80	Very poor
7.55	521	12.4	3.59	291	3.67	76.42	Very poor
7.76	525	11.9	3.36	293	4.3	74.35	Poor
7.80	523	14.5	3.10	292	3.95	74.94	poor
8.05	600	10.3	7.42	300	5.41	101.97	Unsuitable for drinking

The objective behind collecting data related to these parameters is to develop two models: an analytical model based on the calculation of the WQI index, and another model based on machine learning, both allowing the identification of water quality states through the artificial calculation of the WQI. Experimental descriptive data collected over a long period, several years, could help achieve this objective. It is therefore useful to have at least one year to archive data in order to establish a sufficiently complete knowledge base capable of functioning properly. This highlights the need for a knowledge base rich in information, which first requires data collection over a long period, as well as the involvement of an expert.

III.3 Proposed framework

The five steps of the proposed intelligent WQ assessment framework are as follows: (i) Physicochemical WQ sample data collection from the study area, (ii) WQ data is used to

calculate the WQI using the weighted arithmetic average method, (iii) Application of Machine learning models (RF and SVR), (iv) Evaluation and analysis to identify the most suitable model. Figure 3.4 shows details of how the suggested framework was developed.

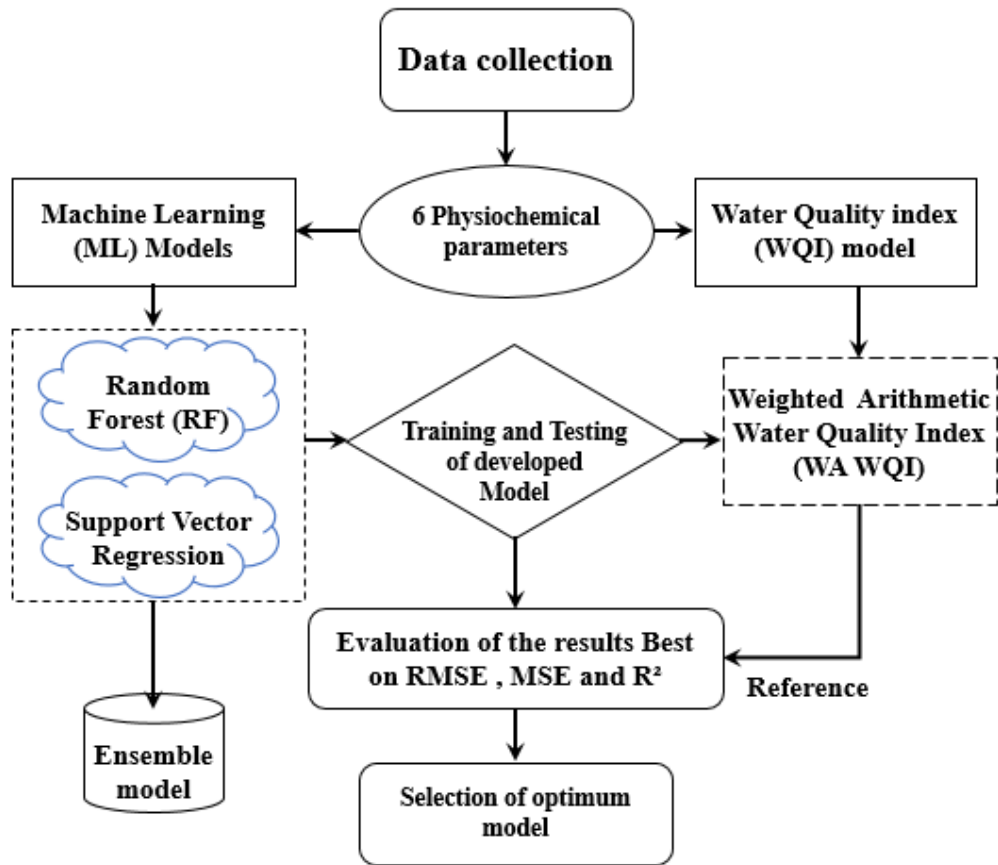


Figure 3.3. Flowchart of the proposed methodology for WQI prediction using ML models.

Developing intelligent models for comprehensive water quality assessment using implemented sensors for data collection is the objective of this work. In order to process large datasets, this advanced system (Figure 3.3) often integrates particular hardware and software components. It is further improved, feature selection and other data preprocessing techniques, as well as classification or prediction algorithms for decision-making. To find the best choice for a water quality assessment, all of these procedures are carried out in our advanced framework for the suggested approach of water quality assessment.

III.4 Results and Analysis

III.4.1 Model Selection

The two ML models RF and SVR are employed to predict the WQI after data preprocessing and to identify patterns and predict future events [37]. The aim of this study is

to assess the effectiveness of these models in predicting the WQI. Both the two models are trained and tested using the WQ dataset, which consists of 800 samples collected from the Tilesdit Dam station in Bouira, Algeria, covering the period from 2016 to 2018.

SVR uses kernel functions to map the input data into high-dimensional feature spaces, and predict water quality indices. Some of the important kernels are Linear, Polynomial, Gaussian RBF and Sigmoid. We have two important hyperparameters: ϵ (epsilon), which is the width of the margin, and C , the regularization parameter that trades off correct classification of training examples against maximum margin (the size of ϵ). Because SVR tries to fit points that are within the margin, support vectors come into play. This optimization aims to maximize the margin while keeping prediction errors at $\leq \epsilon$. This SVR model was designed in Python and fine-tuned for a comparison with the RFR model [38]. The RF model is used in this work to assess and forecast the effects of water quality indicators. A RF model is made up of a huge number of independent decision trees that function as an ensemble. For this regression job, the final forecast is calculated by averaging the forecasts of each individual tree. In this study, Python was used to implement the models [39] [40].

III.4.2 Model Evaluation Metrics

During the study period, both RF and SVR were used to model characteristics of the actual WQI values versus modeled values [41]. The following statistical indicators have been chosen to assess the prediction correctness and forecasting accuracy of those fashions:

➤ **R² (Score):** Indicates the proportion of variance for the WQI that is explained by the input physicochemical parameters. Its value ranges between 0 and 1; the closer it is to 1, the better the model is at understanding changes in water quality. This means that the selected input parameters are the primary factors that account for the variations observed in water quality.

➤ **Mean Squared Error (MSE):** calculates the average squared difference between predicted and actual values of WQI, by squaring the differences (the errors), this metric magnifies large errors, resulting in significantly higher values that may indicate a flaw in the training phase. Its value starts from 0 and increases as the error grows; a lower MSE indicates that the model is more stable and less prone to making large predictive errors.

➤ **Root Mean Squared Error (RMSE):** is the square root of the MSE and measures the average magnitude of errors using the same units as the WQI. It reflects how far the predicted

values deviate from the actual values. The closer the RMSE is to 0, the more accurate the model and the more reliable it becomes for monitoring water quality.

➤ **Mean Absolute Error (MAE):** the Average Value of (MAE) is calculated as the average absolute difference between the predicted and actual values. By ignoring the error's direction to take the absolute value, a low MAE indicates that the model's daily predictions consistently closely match the measured values. MAE is sensitive but not as sensitive as RMSE. When used together, the model's errors are stable and consistent across 800 samples.

III.4.3 Training and Testing results

The main purpose of the dataset is to design and confirm a predictive strategy for WQI prediction, which will improve model generalization by enriching the dataset to more accurately represent current water quality parameters. As seen in Table 3.1, the six parameters in the dataset—Temp, pH, EC, Turb, TDS, and DO—are all derived from direct observations.

The calculated WQI values serve as reference points for training the two machine learning models and evaluating the accuracy of their predictions, namely the SVR and RF models. The next step is to use these WQI values to train and test how well the two selected models predict WQI. This shift moves the study from a traditional analytical method toward machine learning methods that support the forecasting of future WQI values.

In preparation for both models, we divided the dataset into two sub-datasets: a training set 80% (640 Samples) for model building (representative enough) and a test set 20% (160 Samples) for validation, using 5-Fold cross-validation based on WQ parameters, were split into five groups of 128 samples each. Four subsets were used for training and the fifth for internal validation in each iteration. This process of rotation guarantees that each data point is used for training and testing, which minimizes the risk of overfitting and ensures the reliability of the WQI predictions for the Tilesdit Dam. All water quality metrics used on input combinations are standardized using the Min-Max scaler to ensure that the values were within a specified range, typically between 0 and 1 [41]. This cross-validation method helps mitigate overfitting and provides a robust estimate of the model performance on independent datasets. Performance was evaluated using metrics such as MAE, MSE, RMSE, and R^2 . Our simulation investigation employed the following hardware: a personal PC with a 2.11 GHz Intel Core™ i7-8665U processor and 16 GB of RAM. All proposed solutions were implemented using Anaconda Navigator 2.6.3 and Python 3.12 in a Windows 10 (64-bit) operating system environment.

Using GridSearchCV with 5-fold cross-validation, we systematically tuned three important parameters for the SVR prediction method: the penalty factor C (tested values [0.1, 1, 10, 100]), the RBF kernel bandwidth γ (evaluated [0.01, 0.1, 1, 'scale', 'auto']), and the epsilon-tube width ε (explored [0.01, 0.1, 0.5]). As a result, while bad parameter selection may result in underfitting or overfitting issues, appropriate parameter selection in both models can produce outstanding outcomes. To choose the best model for WQI prediction, it is crucial to take into account the chosen performance measures for the two models.

To acquire the training and testing results, we can evaluate the model's results during both steps by analyzing performance metrics between the observed and predicted WQI values of the two models. A thorough comparison of the RF and SVR models' performance is shown in Table 3.6, which also includes the average metric values from the 5-fold cross-validation. To choose the optimal model for prediction, these indicators must be identified during both the training and validation stages. These measures are essential for comprehending prediction error margins, model accuracy, and the percentage of WQI variance that can be accounted for by the water quality parameters.

Table 3.6. Average performance metrics of RF and SVR methods for WQI prediction using 5-Fold cross-validation on training and testing datasets [42] [43] [44].

Model	Training data				Testing data			
	MSE	RMSE	MAE	R ²	MSE	RMSE	MAE	R ²
SVR	0.0013	0.04	0.02	0.999	0.001	0.03	0.02	0.999
RFR	0.27	0.52	0.37	0.997	1.02	1.01	0.78	0.986

The results show that both models perform well, but SVR with GridSearchCV ($C = 500$, $\varepsilon = 0.01$, $\gamma = 0.001$), gives much better predictive accuracy than RFR. For the SVR model, the error values are extremely low in both training and testing data. The training MSE is 0.0013, RMSE is 0.04, and MAE is 0.02, while the testing MSE is 0.001, RMSE is 0.03, and MAE is 0.02. The R^2 value is 0.999 for both training and testing. This means the SVR model explains almost 99.9% of the variation in the target variable. The similarity between training and testing results also shows strong generalization, with no clear sign of overfitting. The RFR model also shows good performance, especially because its R^2 values are high. It achieves an R^2 of 0.997 in training and 0.986 in testing. This means the model explains 99.7% of the variance in the training data and 98.6% in the testing data. Still, its prediction errors are much higher than those of SVR. For example, the testing RMSE of RF is 1.01, compared with only 0.03 for SVR. The testing MAE is also higher, 0.78 for RF compared with 0.02 for SVR.

The difference between training and testing performance is larger in the RF model. Its MSE increases from 0.27 in training to 1.02 in testing, which suggests a slight reduction in predictive stability when applied to unseen test data. This does not indicate severe overfitting, because the testing R^2 remains high, but it shows that RF is less accurate and less stable than SVR.

Figures 3.5, 3.6, 3.7, and 3.8 display the training and testing results for WQI prediction. Plots comparing real and predicted WQI values for the SVR and RF models, respectively, using the training and testing sets are shown in Figures 3.5 and 3.7. The link between the actual and anticipated WQI values for each ward, based on the SVR and RF models, is depicted in scatter plots in Figures 3.6 and 3.8. A thorough comparison of the models' performance across various data subsets is made possible by these figures, which show how effectively the models perform on the target variable.

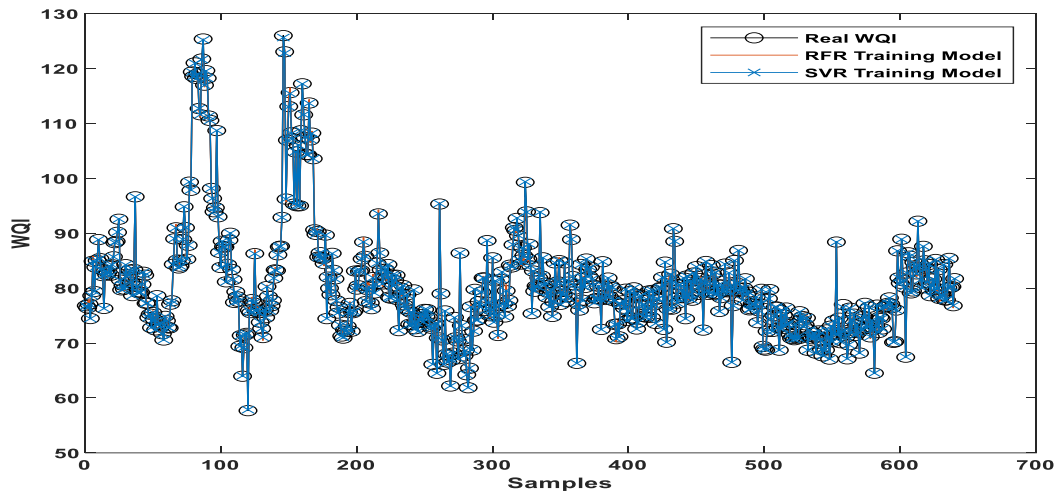


Figure 3.4. WQI values prediction results for SVR and RF models in the training phase.

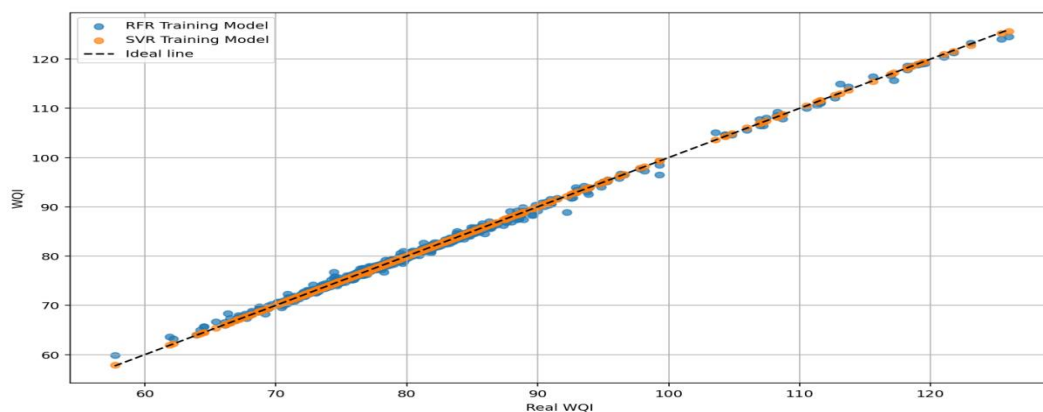


Figure 3.5. Scatter plot of predicted WQI values using SVR and RF models in the training phase.

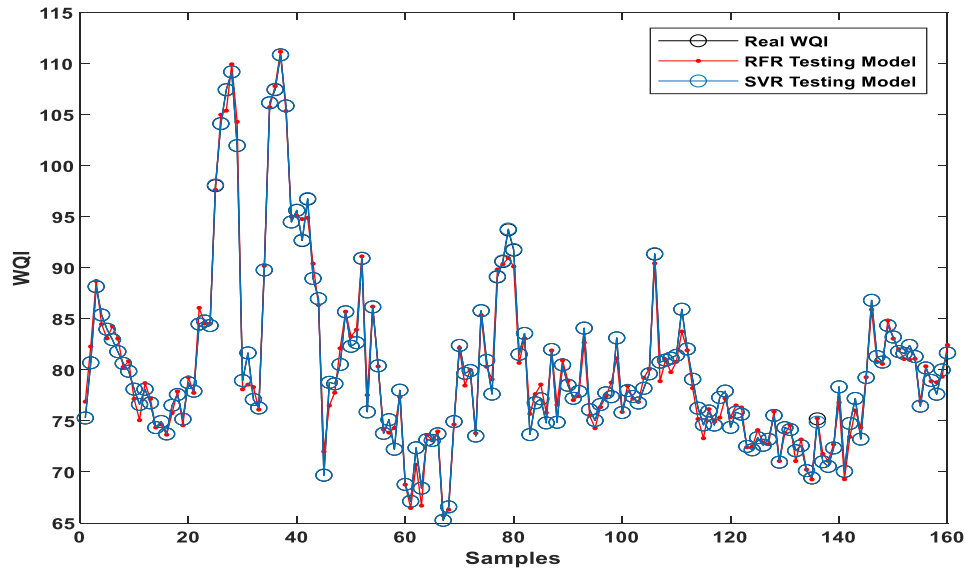


Figure 3.6. WQI values prediction results for SVR and RF models in the testing phase.

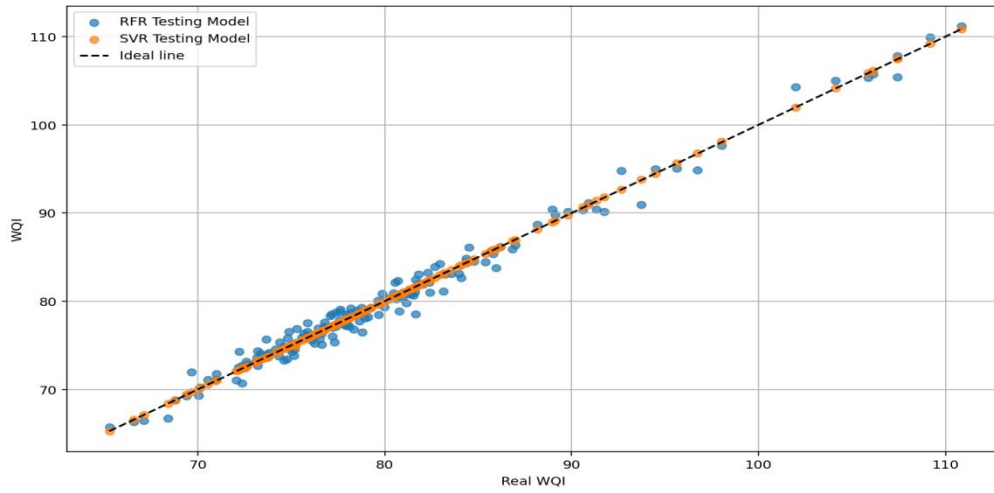


Figure 3.7. Scatter plot of predicted WQI values using SVR and RF models in the testing phase.

Overall, SVR is the best-performing model in this comparison. It provides the lowest prediction errors and the highest R^2 values for both training and testing data. Based on these results, SVR is more suitable for prediction than RF. The near-perfect SVR results should still be checked carefully to make sure there is no data leakage, repeated samples, or preprocessing error, especially if the dataset is small.

III.4.4 Feature selection and importance analysis using RF model

Selecting features is a crucial stage for enhancing both the effectiveness and clarity of a model, particularly because environmental datasets frequently experience multicollinearity and variable significance issues. This study had two main goals in choosing the most impactful features for WQI prediction: first, we wanted to simplify the model by reducing the number of features without losing any information, and second, to improve the predictive

accuracy of the Random Forest model by concentrating on the most important physicochemical parameters. This dual task learning framework using the RF approach structured the model to prioritize the most informative inputs, promoting more stable training dynamics and enhanced interpretability. This section outlines the method used to evaluate feature importance and presents the ranked outcomes applied for additional model training [42].

In developing the dual-task RF model compared to SVR model, this study used a real database of 800 samples collected between 2016 and 2018 from Tilesdit Dam in Bouira, Algeria, including pH, temperature (T°), conductivity (C), turbidity (Turb), total dissolved solids (TD), and dissolved oxygen (DO). The index was initially computed analytically using the Weighted Arithmetic Index approach. These parameters were selected because they are the most frequently monitored at the production station at the study site. For a comprehensive assessment, the dataset was randomly split into training and testing subsets using a hold-out approach, with 80% allocated to training and validation and 20% reserved for testing the WQI prediction performance through metrics such as MAE, RMSE, and R^2 . The data were then normalized using the Min-Max scaler to place all values within a defined range, usually between 0 and 1. For the SVR model, systematic hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation.

WQI prediction using both the full set and several reduced sets of indicators, in order to assess how reducing the number of parameters affects the accuracy of WQI estimates compared with analytical methods. In order to optimize the accuracy of WQI prediction by using a minimal set of water characteristics, the study sought to identify important water parameters and the most dependable models, taking seasonal variations in the water environment into account [45].

Model performance through the learning steps was assessed by comparing the calculated and predicted WQI values using evaluation metrics. Figure 3.9 shows the importance of each input water quality parameter, as determined by the RF method while tables 3.7 and 3.8 presents the performances metric values of RF and SVR methods respectively, with both full and reduced indicator sets.

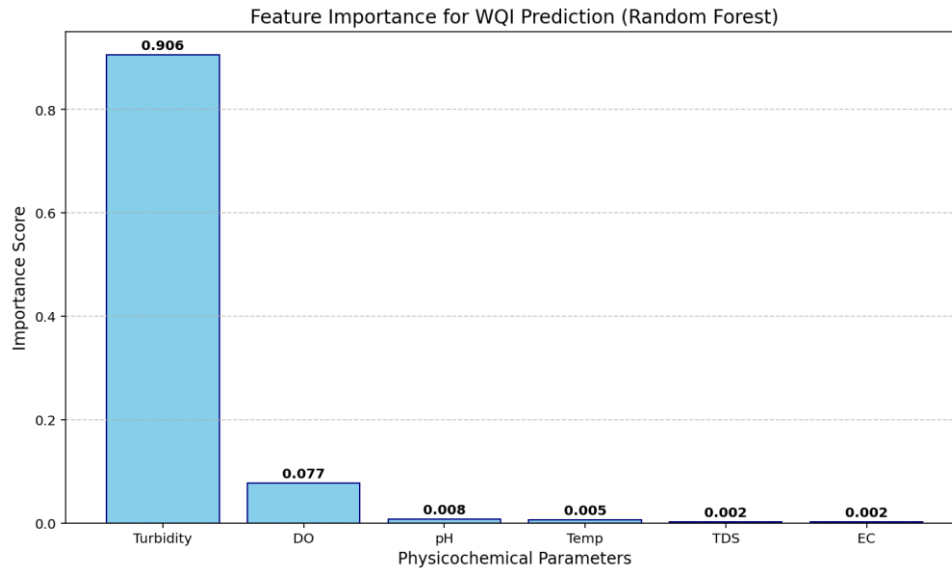


Figure 3.8. Importance of input parameters (RF)

Table 3.7. Performance metrics for RF-based feature selection and WQI prediction (best results are in bold) [44] [45] .

Nb_variables	Variables	R ²	RMSE	MAE
Full indicators	pH, C, T°, Turb, TDS, DO	0.985814	1.013995	0.782286
4	Turb, DO, pH, T°	0.986929	0.973312	0.734582
3	Turb, DO, pH	0.986058	1.005221	0.775226
2	Turb, DO	0.976875	1.294622	0.99573
1	Turb	0.808777	3.722814	2.923929
5	pH, C, T°, TDS, DO	0.451788	6.303409	4.422737

Table 3.8. Performance metrics for RF-based feature selection and SVR based WQI prediction (best results are in bold). [43][44]

Nb_variables	Variables	R ²	RMSE	MAE
Full indicators	pH, C, T°, Turb, TDS, DO	0.999850	0.104293	0.079780
4	Turb, DO, pH, T°	0.999845	0.105898	0.066893
3	Turb, DO, pH	0.995437	0.575059	0.422346
2	Turb, DO	0.982021	1.141528	0.857830
1	Turb	0.858698	3.200182	2.480106
5	pH, C, T°, TDS, DO	0.478043	6.150618	4.195459

Table 3.7 shows that the RF model achieved its best WQI prediction performance when four variables were used: Turb, DO, pH, and T°. This combination produced the highest R² value, 0.986929, and the lowest error values, with RMSE = 0.973312 and MAE = 0.734582. These results are slightly better than the model using all indicators, which recorded R² = 0.985814, RMSE = 1.013995, and MAE = 0.782286. This means that removing less informative variables improved the model slightly and reduced prediction error.

The three-variable subset, Turb, DO, and pH, also gave strong results, with $R^2 = 0.986058$, $RMSE = 1.005221$, and $MAE = 0.775226$. Its performance remains close to the four-variable subset, showing that Turb, DO, and pH are the most influential predictors for WQI estimation. The two-variable model, Turb and DO, still performed well, but its error increased, which indicates that pH and T° add useful information to the model.

The weakest performance appears when Turb is excluded from the selected variables. The five-variable subset, pH, C, T° , TDS, and DO, recorded $R^2 = 0.451788$, $RMSE = 6.303409$, and $MAE = 4.422737$. This sharp decline confirms the dominant role of turbidity in predicting WQI. Even when used alone, Turb achieved $R^2 = 0.808777$, which supports its strong explanatory power.

Table 3.8 shows that SVR produced much higher prediction accuracy than RF when using the same RFR-based feature selection strategy. The full-indicator SVR model achieved the highest R^2 , 0.999850, and the lowest RMSE, 0.104293. The four-variable subset, Turb, DO, pH, and T° , gave almost identical performance, with $R^2 = 0.999845$ and $RMSE = 0.105898$, while producing the lowest MAE, 0.066893. This means that the four-variable subset provides nearly the same predictive accuracy as the full model while using fewer input variables.

The SVR results also confirm the importance of Turb. When Turb was excluded in the five-variable subset, model performance dropped sharply to $R^2 = 0.478043$, $RMSE = 6.150618$, and $MAE = 4.195459$. In contrast, Turb alone produced $R^2 = 0.858698$, which is higher than the single-variable RF result. This confirms that turbidity is the key variable for WQI prediction in this dataset.

Overall, the best reduced feature set is Turb, DO, pH, and T° . It offers the best balance between accuracy and model simplicity. The comparison also shows that SVR is more accurate than RF for WQI prediction, especially when using the full indicators or the four most relevant variables. The RF model provides a compact method for assessing feature importance [43]. The much poorer result obtained with five variables, compared with the full model and the four-variable model, confirms that turbidity plays a dominant role in accurate WQI prediction, as it is the most important input influencing parameter.

Although the RF model yielded lower prediction results compared to the SVR model in the prediction task, it remains valuable for feature selection and feature importance analysis.

Within this dual-task learning framework, the RF approach identifies the most relevant variables, while the SVR model is used for the final WQI prediction.

III.5 Discussion and Practical Implications

The results demonstrate that integrating machine learning techniques establishes a robust framework for environmental water quality monitoring. The system operates through three main components:

III.5.1 Intelligent Data Aggregation

The system functions as an intelligent aggregator, processing four key inputs-Turbidity, Dissolved Oxygen, pH, and Temperature—to estimate a comprehensive WQI using patterns learned via RF and SVR algorithms. This approach reduces reliance on time-consuming laboratory titrations, minimizes human error, and lowers operational costs. It allows field operators to make timely decisions based on reliable, synthesized data.

III.5.2 Robustness and Generalization

Analysis of training and testing results confirms that the system achieves high generalization, capturing the underlying chemical relationships among water quality parameters rather than simply memorizing the data. GridSearchCV effectively optimized the SVR parameters, while the RF algorithm demonstrated superior stability in handling the non-linear noise inherent in sensor measurements.

III.5.3 Practical Real-Time Deployment

Once trained, the models require minimal computational resources, making them suitable for deployment on embedded systems or microcontrollers at the Tilsdit Dam. Integration into a local monitoring station enables real-time alerts whenever the WQI deviates from safe thresholds. This improvement addresses delays associated with traditional manual methods, allowing authorities to respond promptly to potential contamination events and enhance water resource management.

III.6 Conclusion

This chapter establishes the methodological framework for the intelligent prediction of water quality. The SVR and RF techniques were applied to develop WQI modeling in the study area. Presented as a prediction problem, both techniques produced very good simulation results. The real database from the Tilsdit station in Bouira, applied to these methods, included a large dataset collected for six water quality parameters, demonstrating the

effectiveness of both approaches. A rigorous data preprocessing pipeline—including data cleaning via linear interpolation and feature scaling through normalization and standardization—ensured that the raw sensor data was reliable and balanced. Additionally, the detailed presentation of the SVR and RF algorithms provides a clear understanding of the mathematical models selected to capture the non-linear complexities of the reservoir's water quality index. The strategic integration of these machine learning methods, along with K-fold cross-validation and standard evaluation metrics (R^2 , MSE, MAE, and RMSE), establishes a solid foundation. The WQI proved to be a useful tool for decision-making and comparative assessment of water quality over time and across space. Subsequently, the six key parameters were selected to develop the two methods. High correlation coefficient values, with R^2 exceeding 0.98, were obtained for both models during the training and testing phases. The SVR method successfully predicted the WQI with high accuracy. Although the RF model yielded lower prediction results than the SVR model, it remains valuable for feature selection and analysis of variable importance. Within this dual-task framework, the RF approach identifies the most relevant variables, while the SVR model is used for the final WQI prediction.

This research demonstrates that statistical learning-based methods are powerful and reliable for estimating water quality in wetlands and other water bodies. These techniques are recommended for accurate, rapid, and cost-effective water quality assessments across production systems worldwide.

Conclusion

The sustainable management and ecological protection of water resources represent a critical environmental challenge that demands modern, high-reliability engineering solutions. Traditional water quality monitoring frameworks rely heavily on periodic manual sampling and laboratory chemical analyses, which suffer from delayed response times and vulnerability to data gaps caused by hardware malfunctions. To overcome these technological barriers, this thesis has successfully designed and evaluated an intelligent, data-driven framework that bridges Electronic Instrumentation with advanced computational intelligence, taking the Tilesdit Dam in Bouira Province, Algeria, as a vital regional case study.

The systematic execution of this research was carried out across three highly integrated phases based entirely on a customized Python environment:

- First, a comprehensive analysis of environmental context and historical data was established, confirming the importance of Water Quality Indices (WQIs) in reducing complex, multi-dimensional physicochemical datasets into a single, dimensionless, and interpretable metrics.
- Second, a rigorous data preprocessing pipeline was developed. A dataset consisting of 6 key parameters was condition-monitored, utilizing normalization and standardization to ensure stable signal dynamics. The empirical core of the study successfully executed the machine learning models, specifically Support Vector Regression (SVR) and Random Forest (RF). The data was methodically split into 80% for training and 20% for testing, utilizing the k-fold cross-validation technique to predict both the Weighted Arithmetic Water Quality Index (WAWQI).
- Third, a feature importance and selection strategy was executed using Random Forest to determine the most influential inputs.

The numerical results obtained in this study have significant implications for the design of next-generation environmental monitoring stations. The physicochemical processes governing water chemistry are inherently stochastic and non-linear, rendering traditional linear sensor calibration curves insufficient. The machine learning architectures employed in this research effectively captured these complex transformations. The RF model demonstrated exceptional robustness due to its ensemble bootstrap aggregation structure, while the SVR framework achieved high accuracy in minimizing standard telemetry performance metrics.

To evaluate the validity of the two models, SVR and RF, performance metrics utilizing MSE, RMSE, MAE, and R2 were used with k-fold cross-validation. The results indicate that the SVR model offers the highest performance in training and testing phases with R2 superior at 0.99 compared to the results of RF model. According to performance metrics, the SVR model outperformed the RF model during both the training and testing steps, while the RF model also performed effectively. With lower error rates, a higher correlation with observed values, and robust performance on new data, the SVR model outperformed other models in predicting the WQI, making it an accurate and sustainable alternative for traditional methods and a valuable tool for monitoring and evaluating surface water quality in the study area.

In water treatment and dam monitoring, physical probes capable of measuring complex chemical concentrations are expensive, require extensive maintenance, and are prone to chemical drift over time. By deploying the optimized, low-power machine learning models developed in this study, it is possible to mathematically replicate the behaviour of these costly hardware probes. Using input signals from simple, durable physical sensors, the embedded intelligence layer can generate precise real-time estimates of complex parameters and global WAWQI scores. Feature selection and importance analysis substantially reduce structural deployment costs and enable continuous, automated monitoring without human intervention and it can improve the models' predictive ability and offer accurate and well selected parameters for WQI prediction. Advanced methods, such as correlation analysis, data preprocessing, hyperparameter tuning and optimization, and the removal of unnecessary parameters, should be used to evaluate the relative significance of different input WQ parameters in the prediction process.

In order to improve the prediction of the WQI, future research ought to try to extend by using a larger database, including soft sensors for chemical features that are not continually measurable. To improve the prediction model's accuracy, researchers can investigate various architectures, optimization strategies, hybrid approaches, and other innovative deep learning techniques.

This study demonstrates that the evolution of Electronic Instrumentation is deeply dependent on its synergy with artificial intelligence. The smart monitoring framework established for the Tiledit Dam shifts the paradigm from passive data logging to active, real-time prediction within Embedded Systems. This scalable, robust, and cost-effective blueprint can be readily implemented across various water infrastructures in Algeria, ensuring the intelligent preservation of our most vital natural resource.

Bibliography

- [1] « Prediction of Water Quality Index (WQI) Using Machine Learning ». *International Journal of Environmental Science and Development*, vol. 16, no 1, 2025. DOI.org (Crossref), <https://doi.org/10.18178/ijesd.2025.16.1.1507>.
- [2] Choudhary, R., Kumar, A., C., P. et al. Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence. *Sci Rep* 15, 31139 (2025). <https://doi.org/10.1038/s41598-025-09463-4>
- [3] Bouslah, Soraya, et al. « Water quality index assessment of Koudiat Medouar Reservoir, northeast Algeria using weighted arithmetic index method ». *Journal of Water and Land Development*, vol. 35, no 1, décembre 2017, p. 221 28. DOI.org (Crossref), <https://doi.org/10.1515/jwld-2017-0087>.
- [4] S. Sarwar, I. Ahmmed, S. Mustari et M. R. Shaibur, «Use of Weighted Arithmetic Water Quality Index (WAWQI) to Determine the Suitability of Groundwater of Chaugachcha and Manirampur Upazila, Jashore, Bangladesh,» *Environmental and Biological Research*, vol. 2, n° %12, pp. 37-48, 2020
- [5] L. Tang, J. Ma, Q. Zhang, X. Gu, G. Ma et J. Zhou, «Evaluation of surface water quality using water quality index (WQI) and human health risk assessment (HHRA) in a typical karst basin, Southwest China,» *Ecological Indicators*, vol. 155, p. 110996, 2023.
- [6] T. D. BANDA et M. KUMARASAMY, «AGGREGATION TECHNIQUES APPLIED IN WATER QUALITY INDICES (WQIS),» *Poll Res.*, vol. 39, n° %12, p. 401, 2020.
- [7] R. Guerrou et R. Abdellaoui, Contribution to the assessment of the quality of some groundwater wells in the Ouargla region using the Water Quality Index (WQI), *Ouargla: University of Kasdi Merbah–Ouargla-*, 2025, p. 7.
- [8] T. Poonam, T. Bhattacharya et S. Chakraborty, «Water Quality Indices – Important Tools for Water Quality Assessment: A Review,» *International Journal of Advances in Chemistry*, vol. 1, n° %11, pp. 15-28, 2013.
- [9] S. Chidiac, P. El Najjar, N. Ouaini, Y. El Rayess et D. El Azzi, «A comprehensive review of water quality indices (WQIs):history, models, attempts and perspectives,» *Rev Environ Sci Biotechnol*, vol. 22, n° %12, p. 352, 2023.
- [10] S. Chidiac, P. El Najjar, N. Ouaini, Y. El Rayess et D. El Azzi, «A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives,» *Rev Environ Sci Biotechnol*, vol. 22, pp. 349-395, 2023.
- [11] M. G. Uddin, S. Nash et A. I. Olbert, «A review of water quality index models and their use for assessing surface water quality,» *Ecological Indicators*, vol. 122, p. 107218, 2021.
- [12] G. Darvishi, F. G. Kootenaei, M. Ramezani, E. Lotfi et H. Asgharnia, «Comparative Investigation of River Water Quality by OWQI, NSFWQI and Wilcox Indexes (Case study: the Talar River – IRAN),» *Archives of Environmental Protection*, vol. 42, n° %11, p. 42, 2016.
- [13] «A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives,» *Reviews in Environmental Science and Bio/Technology*, vol. 22, p. 355, 2023.
- [14] T. Shweta, S. Bhavtosh, S. Prashant et D. Rajendra, «Water Quality Assessment in Terms of Water Quality Index,» *American Journal of Water Resources*, vol. 1, n° %13, p. 36, 2020.
- [15] V. A. Călmuc, M. Călmuc, M. C. Țopa, M. Timofti, C. Iticescu et L. P. Georgescu, «VARIOUS METHODS FOR CALCULATING THE WATER QUALITY INDEX,» *Annals of the "Dunărea de Jos" University of Galati. Fascicle II, Mathematics, Physics, Theoretical Mechanics*, vol. 41, n° %12, p. 173, 2018.

- [16] M. D. Noori, «Comparative analysis of weighted arithmetic and CCME Water Quality Index estimation methods, accuracy and representation,» IOP Conf. Series: Materials Science and Engineering, vol. 737, n° %11, p. 012174, 2020.
- [17] T. Shweta, S. Bhavtosh, S. Prashant et D. Rajendra, «Water Quality Assessment in Terms of Water Quality Index,» American Journal of Water Resources, vol. 1, n° %13, p. 35, 2013.
- [18] S. BRIK et A. DECHOUCHA, Contribution à l'étude et l'évaluation des indices de qualité des eaux par l'élaboration d'un outil décisionnel intelligent, M'sila : Université Mohamed Boudiaf - M'sila , 2021.
- [19] S. Ahmed Amin, «Contribution to the study of dams' water quality in Algeria,» Université Mohamed Boudiaf - M'sila, M'sila, 2021.
- [20] . R. Guerrou et R. Abdellaoui , Contribution to the assessment of the quality of some groundwater wells in the Ouargla region using the Water Quality Index (WQI), Ouargla: University of Kasdi Merbah–Ouargla-, 2025.
- [21] S. Hamouche et K. Sebai, Apport de la télédétection et des SIG pour le suivi de la qualité physico chimique des eaux de surface.Cas de Barrage de Beni Haroun (W. de Mila), M'SILA: UNIVERSITE MOHAMED BOUDIAF - M'SILA, 2021.
- [22] A. BENFERDIA et M. . A. KORICHI, «Evalauton de la qualité des eaux souterraine par outil SIG . cas de la partie aval du bassin versant d'oued Djendjen (Jijel),» UNIVERSITE KASDI MERBAH – OUARGLA, OUARGLA, 2021.
- [23] Z. Z. Aljanabi, A.-H. M. Jawad Al-Obaidy et F. M. Hassan, «A brief review of water quality indices and their applications,» IOP Conference Series: Earth and Environmental Science, vol. 779, n° %11, p. 012088, 2021.
- [24] S. BRIK et A. DECHOUCHA, Contribution à l'étude et l'évaluation des indices de qualité des eaux par l'élaboration d'un outil décisionnel intelligent, M'sila : Université Mohamed Boudiaf - M'sila , 2021.
- [25] N. El Hanafi, "Conception et dimensionnement de l'extension de la station de traitement d'eau potable (100 000 m3/j) Tilesdit/W de Bouira," Master's thesis, Dept. of Process Engineering, University of Bouira, Bouira, Algeria, 2022
- [26] Z. Chen, P. Jiang, J. Liu, S. Zheng, Z. Shan, and Z. Li, "An adaptive data cleaning framework: a case study of the water quality monitoring system in China," Hydrological Sciences Journal, vol. 67, no. 7, pp. 1114-1129, May 2022. doi: 10.1080/02626667.2022.2060106
- [27] R. Li, "Normalization vs. Standardization: Key Differences Explained," DataCamp, 2023. [Online]. Available: <https://www.datacamp.com/blog/normalization-vs-standardization>. [Accessed: May 10, 2026].
- [28] S. J. Al-Rubaie and A. S. Al-Hussaini, "Statistical analysis and correlation of water quality parameters in dam reservoirs," Journal of Hydrology, 2021
- [29] P. Sahu, S. N. Londhe, and P. S. Kulkarni, "Prediction of water quality parameters using support vector regression," Innovative Infrastructure Solutions, vol. 8, no. 229, Aug. 2023. doi: 10.1007/s41062-023-01195-2.
- [30] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
- [31] A. Al-Nahid, M. A. Hasan, and others, "Water quality prediction using random forest and other machine learning techniques," Journal of Environmental Management

- [32] L. Latifoğlu and A. Doğan, "Multi-source machine learning models for peak prediction with LASSO feature selection: a Susquehanna Basin case study," *Hydrological Sciences Journal*, vol. 71, no. 4, pp. 1-16, Apr. 2026. doi: 10.1080/02626667.2026.2653055
- [33] S. C. Mukhopadhyay and C. P. Kurup, *Advanced Machine Learning for Integrated Instrumentation and Sensing Systems*, 1st ed. Cham, Switzerland: Springer Nature, 2023
- [34] M. Ehteram, A. N. Ahmed, M. Sherif, and A. El-Shafie, "An advanced deep learning model for predicting water quality index," *Ecological Indicators*, vol. 160, p. 111806, 2024
- [35] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed. Springer, 2021.
- [36] F. M. Kizar, «A comparison between weighted arithmetic and Canadian methods for a drinking water quality index at selected locations in shatt al-kufa,» *IOP Conference Series: Materials Science and Engineering*, vol. 433, n° %11, p. 6, 2018
- [37] R. K. Makumbura, L. Mampitiya, N. Rathnayake, D. Meddage, S. Henna, T. L. Dang, Y. Hoshino et U. Rathnayake, «Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature,» *Results in Engineering*, vol. 23, p. 102831, Septembre 2024
- [38] A. Sethi, «Support Vector Regression (SVR) Simplified & How To Tutorial,» *Analytics Vidhya*, 29 03 2020. [En ligne]. Available: <https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>. [Accès le 29 04 2026].
- [39] J. Sun, X. Guan, X. Sun, X. Cao, Y. Tan et J. Liao, «Water quality prediction and carbon reduction mechanisms in wastewater treatment in Northweat cities using Random Forest Regression model,» *Scientific Reports*, vol. 14, n° %11, p. 31525, 28 December 2024.
- [40] N. Boukelkal, S. Rahal, R. Rebhi et M. Hamadache, «QSPR for the prediction of critical micelle concentration of different classes of surfactants using machine learning algorithms,» *Journal of Molecular Graphics and Modelling*, vol. 129, p. 108757, 06 2024.
- [41] B. Sakaa, A. Elbeltagi, S. Boudibi, H. Chaffaï, A. R. M. T. Islam, L. C. Kulimushi, P. Choudhari, A. Hani, Y. Brouziyne et Y. J. Wong, «Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin,» *Environmental Science and Pollution Research*, vol. 29, n° %132, pp. 48491-48508, 07 2022.
- [42] A. A. Bataineh, B. Vamsi et S. A. Smith, «A Hybrid Machine Learning Framework for Water Quality Index Prediction Using Feature-Based Neural Network Initialization,» *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 4887-4905, 2026.
- [43] Habiba BAKACHE et al., Prediction of weighted arithmetic water quality index for surface water quality using Support Vector Machines and Random Forest approaches, *First National Conference on Measurement and Sensors (FNCMS'2026)*, 05-06 May 2026, M'sila –ALGERIA
- [44] Kaouther FOUDILI et al., A Dual-Task Learning Framework for Enhanced Water Quality Index Modeling Using Random Forest-Based Feature Selection and Prediction, *First National Conference on Measurement and Sensors (FNCMS'2026)*, 05-06 May 2026, M'sila -ALGERIA
- [45] Y. Mo, J. Xu, C. Liu, J. Wu et D. Chen, «Assessment and prediction of Water Quality Index (WQI) by seasonal key water parameters in a coastal city: application of machine learning models,» *Environmental Monitoring and Assessment*, vol. 196, n° %111, p. 1008, 11 2024.

ملخص

يُعد مراقبة جودة المياه أمراً أساسياً لإدارة وحماية موارد المياه السطحية. تبحث هذه الدراسة في تقييم وتنبؤ جودة المياه باستخدام كل من الأساليب التحليلية وتقنيات التعلم الآلي. تقوم الأساليب التحليلية لحساب مؤشر جودة المياه (WQI) بتجميع عدة معايير لجودة المياه في قيمة واحدة تمثل الجودة الإجمالية للمياه. يُعد تبني طرق موثوقة للتعلم الآلي في التنبؤ بمؤشر جودة المياه أمراً حيوياً لتحقيق نتائج دقيقة من خلال الاستفادة من المعالجة الحاسوبية لتقريب المؤشر بناءً على متغيرات إدخال معقدة ومجمعة. تُقيم هذه الدراسة فعالية كل من الانحدار بواسطة المتجهات الداعمة (SVR) والانحدار بالغابات العشوائية (RFR) في التنبؤ بمؤشر جودة المياه للمياه السطحية. تم حساب مؤشر جودة المياه باستخدام طريقة الحساب الحسابية الموزونة لتحديد فئات الجودة الأساسية وتقييم جودة المياه السطحية عند سد تيلسديت في البويرة، الجزائر، استناداً إلى مجموعة بيانات تتضمن ستة معايير فيزيوكيميائية رئيسية: الرقم الهيدروجيني، درجة الحرارة، التوصيلية الكهربائية، العكارة، المواد الصلبة الذائبة الكلية، والأكسجين المذاب. تم تقييم أداء نماذج التنبؤ باستخدام عدة مقاييس إحصائية. أظهرت تقييمات الأداء أن نموذج SVR قدم دقة فائقة، مع قيم R^2 تتجاوز 0.999 وأخطاء تنبؤ طفيفة. وأشارت تحليل أهمية المتغيرات باستخدام نموذج RFR إلى أن العكارة كانت العامل الأكثر تأثيراً في التنبؤ الدقيق بمؤشر جودة المياه. نجح نموذج SVR في التنبؤ بالمؤشر بدقة عالية. على الرغم من أن نموذج RFR أظهر نتائج تنبؤ أقل من SVR، إلا أنه يظل ذا قيمة لاختبار المتغيرات وتحديد أهم العوامل ذات الصلة. ضمن هذا الإطار ذو المهام المزدوجة، يحدد نموذج RFR أهمية المتغيرات، بينما يقوم نموذج SVR بالتنبؤ النهائي لمؤشر جودة المياه. يتيح الإطار المقترح مراقبة ذكية وجودة المياه في الوقت الفعلي مع تقليل الاعتماد على التحاليل المخبرية، موفرًا أداة فعالة من حيث التكلفة، سريعة وموثوقة لاتخاذ القرارات البيئية وإدارة الموارد.

الكلمات المفتاحية: المياه السطحية، مؤشر جودة المياه (WQI)، التعلم الآلي، الانحدار

Abstract:

Water quality monitoring is essential for managing and protecting surface water resources. This study investigates the assessment and prediction of water quality using both analytical and machine learning approaches. Analytical methods for calculating the Water Quality Index (WQI) aggregate multiple water quality parameters into a single value representing overall water quality. Adopting reliable machine learning methods for predicting the WQI is crucial for achieving accurate results by leveraging computational processing to approximate the WQI from complex, combined input variables. This study evaluates the efficacy of Support Vector Regression (SVR) and Random Forest Regression (RFR) in predicting the WQI of surface water. The WQI was calculated using the Weighted Arithmetic Method to establish baseline water quality classes and assess the surface water quality at Tilesdit Dam in Bouira, Algeria, based on a dataset of six key physicochemical parameters: pH, temperature, electrical conductivity, turbidity, total dissolved solids, and dissolved oxygen. The performance of the prediction models was evaluated using multiple statistical metrics. Performance evaluation demonstrated that SVR provided superior accuracy, with R^2 values exceeding 0.999 and minimal prediction errors. Feature importance analysis using the RFR model indicated that turbidity was the most influential parameter for accurate WQI prediction. SVR successfully predicted the WQI with high accuracy. Although RFR yielded lower prediction results than SVR, it remains valuable for feature selection and identifying the most relevant variables. Within this dual-task framework, RFR determines variable importance, while SVR performs the final WQI prediction. The proposed framework enables intelligent, real-time water quality monitoring with reduced reliance on laboratory analyses, providing a cost-effective, rapid, and reliable tool for environmental decision-making and management.

Keywords: Surface water, Water Quality Index (WQI), Machines learning, Regression

Résumé :

La surveillance de la qualité de l'eau est essentielle pour la gestion et la protection des ressources en eau de surface. Cette étude examine l'évaluation et la prédiction de la qualité de l'eau en utilisant à la fois des méthodes analytiques et des approches d'apprentissage automatique. Les méthodes analytiques de calcul de l'Indice de Qualité de l'Eau (IQE) regroupent plusieurs paramètres de qualité de l'eau en une seule valeur représentant la qualité globale de l'eau. L'adoption de méthodes fiables d'apprentissage automatique pour prédire l'IQE est cruciale afin d'obtenir des résultats précis en exploitant le traitement computationnel pour approximer l'IQE à partir de variables d'entrée complexes et combinées. Cette étude évalue l'efficacité de la Régression par Vecteurs de Support (SVR) et de la Régression par Forêt Aléatoire (RFR) dans la prédiction de l'IQE de l'eau de surface. L'IQE a été calculé en utilisant la Méthode Arithmétique Pondérée afin d'établir des classes de qualité de l'eau de référence et d'évaluer la qualité de l'eau de surface au barrage de Tilesdit à Bouira, Algérie, sur la base d'un ensemble de données comprenant six paramètres physico-chimiques clés : pH, température, conductivité électrique, turbidité, solides dissous totaux et oxygène dissous. La performance des modèles de prédiction a été évaluée à l'aide de plusieurs indicateurs statistiques. L'évaluation des performances a montré que la SVR offrait une précision supérieure, avec des valeurs de R^2 dépassant 0,999 et des erreurs de prédiction minimales. L'analyse de l'importance des variables à l'aide du modèle RFR a indiqué que la turbidité était le paramètre le plus influent pour une prédiction précise de l'IQE. La SVR a permis de prédire l'IQE avec une grande précision. Bien que la RFR ait fourni des résultats de prédiction inférieurs à ceux de la SVR, elle reste utile pour la sélection des variables et l'identification des paramètres les plus pertinents. Dans ce cadre à double tâche, la RFR détermine l'importance des variables, tandis que la SVR effectue la prédiction finale de l'IQE. Le cadre proposé permet une surveillance intelligente et en temps réel de la qualité de l'eau, avec une dépendance réduite aux analyses en laboratoire, offrant un outil rapide, fiable et économique pour la prise de décision et la gestion environnementale.

Mots-clés : Eaux de surface, Indice de qualité de l'eau (IQE), Apprentissage Automatique, Régression