20CS7551: B. Tech Mini Project – II (Review 1)
Batch No: 11

WATER QUALITY ASSESSMENT

Enhancing Water Quality Assessment through Ensemble Techniques in Sentinel2 Remote Sensing Imagery Analysis

Guide:
Mr. D. Suresh Babu,
Assistant Professor.

Team members:
G. Snehitha (208W1A0522)
P. Keerthi (208W1A0538)

CONTENTS

ABSTRACT

Remote sensing imagery has proven to be a valuable tool for monitoring and assessing water quality in various aquatic environments. This study focuses on the enhancement of water quality assessment through the utilization of ensemble techniques applied to Sentinel-2 satellite imagery data. After atmospheric correction, relevant parameters that include chlorophyll, pH, dissolved oxygen, salinity, turbidity, dissolved organic matter and suspended matter were extracted from Sentinel-2 band radiance values using established research formulae, enabling the characterization of water quality indicators. In this project, we propose the application of gradient boosting algorithm as an ensemble technique to improve the accuracy and robustness of water quality assessment. Gradient boosting has demonstrated success in handling complex datasets and capturing nonlinear relationships, making it a suitable candidate for enhancing the analysis of remote sensing data. The proposed ensemble approach is anticipated to demonstrate improved accuracy in classifying water quality as "good", "bad", or "needs treatment" providing an actionable context to the results compared to individual models. This advancement has the potential to contribute significantly to the field of water quality monitoring, aiding environmental management and policy decisions that rely on accurate and timely information about water resources.

Keywords: Water Quality Assessment, Feature Extraction, Gradient Boosting classifier, Sentinel2 remote sensing Analysis.

AIM

To improve the accuracy of water quality assessment by employing ensemble techniques on Sentinel-2 remote sensing imagery data. This will enable precise classification of water quality into distinct categories, facilitating more informed environmental decision-making and resource conservation.

MOTIVATION

- The motivation for this project arises from the increasing importance of water quality assessment in maintaining the ecological balance and ensuring the health and safety of communities.
- **Traditional assessment methods** may lack the precision and efficiency required for accurate decision-making.
- By harnessing the power of **ensemble techniques** and leveraging **Sentinel-2 remote sensing imagery data**, this project seeks to provide a more sophisticated and reliable means of evaluating water quality.
- This advancement could lead to more proactive environmental management, improved public health, and a sustainable use of water resources.

RESEARCH QUESTIONS

1. How does the use of Sentinel-2 remote sensing imagery data enhance the accuracy and reliability of water quality assessment compared to traditional sampling and testing methods?
2. What are the main factors influencing water quality in each of the lakes?
3. How accurately can water quality be classified into "good", "poor", and "needs treatment" categories using ensemble techniques in remote sensing imagery data for the selected lakes?
4. Can the insights gained from this study contribute to better management strategies and policies aimed at improving water quality and preserving the ecological health of Tappar, Shinai, and Hamisar lakes?
5. Can the methodology developed in this project be extended to assess other environmental parameters, such as sedimentation rates or algal blooms, in Tappar, Shinai, and Hamisar lakes?

TITLE JUSTIFICATION

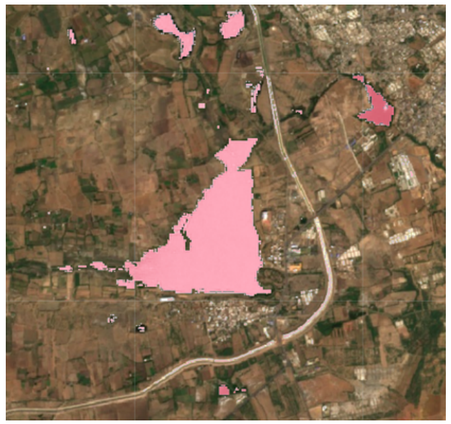
- The justification for the title **"Enhancing Water Quality Assessment through Ensemble Techniques in Sentinel2 Remote Sensing Imagery Analysis"** is rooted in the need for improved water quality assessment methods.
- Traditional approaches might not capture the complexity and nuances of water quality patterns. By emphasizing the use of the **gradient boosting algorithm,** the title highlights an advanced machine learning technique known for its ability to handle complex relationships and improve predictive accuracy.
- Moreover, the utilization of **Sentinel-2 remote sensing imagery data** further adds value by offering a comprehensive and data-rich source for environmental monitoring.
- Thus, the title aptly underscores the project's focus on refining water quality assessment through the amalgamation of cutting-edge algorithms and sophisticated data sources.

OBJECTIVES

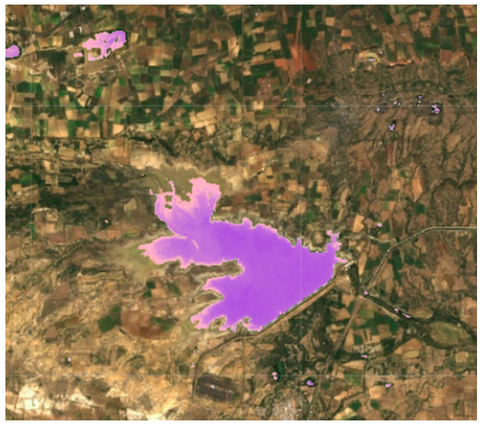
1. To collect and preprocess Sentinel-2 imagery data for water quality assessment
2. To extract features from satellite data that influence water quality.
3. To implement and optimize the Gradient Boosting classifier algorithm for classification.
4. To evaluate and validate the model's accuracy and reliability.
5. To derive insights and recommendations for effective water quality management.

SCOPE

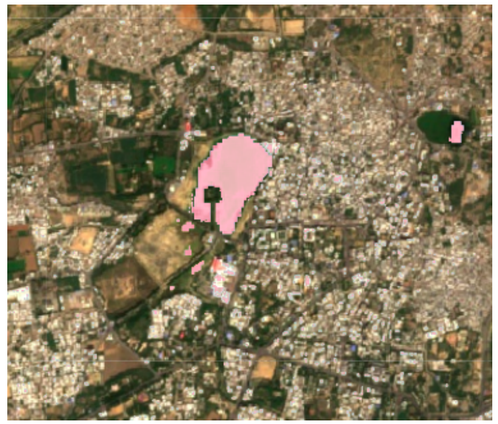
1. The Scope of the project is now limited to three lakes Tappar lake, Shinai lake and Hamisar lake.
2. The parameters for assessing water quality are limited to the number 8 - chlorophyll, pH, Land Surface Temperature, dissolved oxygen, salinity, turbidity, dissolved organic matter.



Shinai lake



Tappar lake



Hamisar lake

LITERATURE SURVEY[1]

Title: Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study

Journal Details: Water Research,Volume 38, Issue 18,2004,ScienceDirect

Dataset: A total of eight monitoring sites were selected for river quality assessment. These sites are identified as Neemsar (Site-1), Bhatpur (Site-2), Gaughat (Site-3), Mid-Lucknow (Site-4), Pipraghat (Site-5), Gangaganj (Site-6), downstream of Sultanpur (Site-7), and downstream of Jaunpur (Site-8).

Description:

This case study focuses on the application of multivariate statistical techniques to assess temporal and spatial variations in water quality data obtained during the monitoring of the Gomti River in Northern India. The study spans a 5-year period from 1994 to 1998, during which water quality data for **24 parameters was collected at eight different sites representing low, moderate, and highly polluted regions of the river**. The complex dataset, comprising 17,790 observations, is analyzed using various **multivariate techniques**, including **cluster analysis (CA), factor analysis/principal component analysis (FA/PCA), and discriminant analysis (DA).** The objective is to uncover patterns, group parameters, and identify **key indicators that contribute to variations in water quality.**

Advantages:

- Effective grouping of sampling sites using Cluster Analysis (CA).
- Identification of key factors explaining data variance via Factor Analysis/Principal Component Analysis (FA/PCA),Significant data reduction for both temporal and spatial analysis with Discriminant Analysis (DA).
- Enhanced understanding of complex water quality data.

Disadvantages:

- Complex data analysis requiring statistical expertise.

LITERATURE SURVEY[2]

Title: A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques
Journal Details: Sensors, Volume-16, Article-1298,2016,MDPI
Dataset: -
Description:
This project focuses on the utilization of remotely sensed data to enhance the effectiveness of monitoring waterbodies. It acknowledges the widespread use of remote sensing techniques to measure key qualitative parameters in water, including **suspended sediments, CDOM, chlorophyll-a, and various pollutants**. The project explores the **vast array of sensors deployed on satellites and aircraft**, designed to capture radiation reflected from the water's surface at different wavelengths. In particular, the project reviews and compiles essential properties such as **spectral, spatial, and temporal characteristics of commonly employed spaceborne and airborne sensors to serve as a valuable guide for sensor selection**. Additionally, the research delves into the examination of prevalent approaches and sensors used for the evaluation and quantification of eleven crucial water quality parameters. These parameters encompass **chlorophyll-a, CDOM, Secchi disk depth, turbidity, TSS, water temperature, total phosphorus, sea surface salinity, dissolved oxygen, BOD, and COD**.
Advantages:

- Diverse sensor options available on satellites and aircraft platforms.
- The project investigates common approaches and sensors used for evaluating and quantifying eleven critical water quality parameters, contributing to a better understanding of water quality dynamics.

Disadvantages:

- Validation through ground measurements adds complexity and cost.

LITERATURE SURVEY[3]

Title: Assessment of Surface Water Quality Using Water Quality Index and Discriminant Analysis Method

Journal Details: Water,VOLUME 15,Article 680,2023,MDPI

Dataset: This dataset is from the Koudiat Medouar watershed in northeastern Algeria, spanning 590 km2, and covers a two-year period (2019-2020).The region exhibits a semi-arid climate with distinctive seasonal variations in temperature and precipitation.

Description:

This study specifically focuses on evaluating surface water quality for drinking and irrigation purposes using the Water Quality Index (WQI) and Irrigation Water Quality Index (IWQI) based on nine hydrochemical parameters. Discriminant analysis (DA) is employed to identify the key variables responsible for spatial variations. Findings indicate that the surface water quality for drinking is generally poor and very poor according to the WQI values, while the IWQI values suggest that the water is acceptable for irrigation with some restrictions for salinity-sensitive plants. pH, potassium, chloride, sulfate, and bicarbonate are identified as significant parameters contributing to spatial variations in surface water quality. The study offers valuable insights for decision-makers in addressing water quality management and protection challenges.

Advantages:

- Use of internationally recognized indexes (WQI and IWQI).
- Weighted parameter evaluation for precision.

Disadvantages:

- Data collection limited to a two-year period.
- Oversimplification of water quality classification.

LITERATURE SURVEY[4]

Title: A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques

Journal Details: Chemometrics and Intelligent Laboratory Systems,volume 214,2021,ScienceDirect

Dataset: Tilesdit dam in Bouira, Algeria, collected over three years (2016–2018)

Description:

This project focuses on enhancing water quality classification through the integration of advanced machine and deep learning techniques. Specifically, **Long Short Term Memory Recurrent Neural Networks (LSTM RNNs) are employed to construct an intelligent model for classifying drinking water quality.** Feature extraction methods, including **Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA)**, are used to reduce data dimensionality and enhance classifier performance.Evaluation methods include various cross-validation techniques and out-of-sample tests. The integration of LSTM RNNs with LDA and ICA demonstrates an impressive accuracy of 91% using the Random-Holdout technique. Comparisons are made with Support Vector Machines (SVMs) models, highlighting the effectiveness of deep learning in water quality classification.

Advantages:

- Enhanced water quality classification accuracy (91%)

Disadvantages:

- It relies on a specific deep learning model, LSTM RNNs, which can be computationally intensive and may require significant computational resources for training and deployment.

LITERATURE SURVEY[5]

Title: Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil

Journal Details: SN Applied Sciences,Volume 3,2021,SpringerLink

Dataset: The Mirim Lagoon hydrographic basin in southern Brazil, comprising 154 water samples collected during 22 sampling campaigns at 7 monitoring points between 2015 and 2017.

Description:

This project centers on the development of a new Water Quality Index (WQI) equation for Mirim Lagoon using **Multiple Linear Regression (MLR)** as the statistical method.The objective was to create a simplified WQI equation by considering only three key variables: **phosphorus, dissolved oxygen (DO), and thermotolerant coliforms.** The MLR model's performance was assessed with an R2 coefficient, which revealed that it could explain 72.8% of the data variability. Additionally, t**he study employed a paired t-test to confirm that the new WQI did not significantly differ from the original WQI,** providing validation for the simplified equation. Using this multiple linear regression (MLR) methodology, also obtained results with an **accuracy of 85%** from a new WQI that can be compared to the original WQI, demonstrating that the new equation modeled by them can be used to predict and monitor the quality of the waters of this river.

Advantages:

- Cost reduction in monitoring and ease of communication with the simplified index.

Disadvantages:

- Oversimplifying water quality assessment by reducing the number of variables, which may not capture all relevant parameters accurately.

LITERATURE SURVEY[6]

Title: Remote Sensing Techniques to Assess Water Quality

Journal Details: Photogrammetric Engineering & Remote Sensing,volume 69,2003,ResearchGate

Dataset: -

Description:

This project centers on the utilization of remote sensing techniques for monitoring and evaluating various water quality parameters in surface water bodies. It emphasizes that substances in surface water can alter backscattering characteristics, and **remote sensing relies on measuring these spectral signature changes and connecting them to water quality parameters through empirical or physically based models.** The selection of the optimal wavelength for measuring a parameter depends on the substance, its concentration, and sensor characteristics. The project addresses major factors affecting water quality, including suspended sediments, algae, chemicals, dissolved organic matter, thermal releases, aquatic plants, pathogens, and oils, and how these factors influence energy spectra in reflected solar and thermal radiation. It discusses the development of remote sensing techniques, with a focus on both empirical models, where statistical relationships are established between spectral properties and water quality parameters, and physically based models that leverage the optical properties of water to estimate parameters like suspended sediments.

Advantages:

- Spatial and temporal monitoring capabilities.
- Focus on key pollutants affecting water bodies.

Disadvantages:

- Limitations of empirical models for specific conditions.
- Indirect inference for certain water quality parameters.

LITERATURE SURVEY[7]

Title: Remote sensing of water quality in an Australian tropical freshwater impoundment using matrix inversion and MERIS images

Journal Details: Remote Sensing of Environment,Volume 115, Issue 9,2011,ScienceDirect

Dataset: Burdekin Falls Dam in Northern Australia and measurements taken using RAMSES spectroradiometers

Description:

This study aimed to adapt and enhance semi-analytical inversion techniques for remote sensing of water quality **parameters (chlorophyll a, tripton, and CDOM)** in Australian tropical and subtropical water bodies, specifically focusing on Burdekin Falls Dam. **The Matrix Inversion Method (MIM) with a semi-analytic mode**l was applied to MERIS images, addressing atmospheric correction challenges. The study compared different weighting schemes and found that overdetermined systems of equations, rather than exact solutions, significantly improved the accuracy and precision of retrieved water quality parameters. This enhanced accuracy has implications for improving water resource management. Additionally, the study validated the MIM approach against field observations and examined the impact of atmospheric correction errors, suggesting reasonable results for Australian water bodies.

Advantages:

- By comparing different weighting schemes, the research identifies the best-performing schemes for estimating chlorophyll-a, tripton, and colored dissolved organic matter (CDOM) concentrations

Disadvantages:

- Small Sample Size

LITERATURE SURVEY[8]

Title: A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges

Journal Details: Remote Sensing,VOLUME 14,ARTICLE 1770,2022,MDPI

Dataset: **Multispectral data like Landsat, Sentinel-2, and SPOT satellites**, suitable for empirical modeling. **Hyperspectral data from satellites like Hyperion and HIS**.While n**on-satellite sources like UAV-based systems** provide flexibility but may have cost and coverage limitations for large water areas.

Description:

This project discusses the application of remote sensing for water quality retrieval, emphasizing data sources and retrieval modes. It covers specific retrieval algorithms for various water quality variables such as total suspended matter, chlorophyll-a, colored dissolved organic matter, chemical oxygen demand, total nitrogen, and total phosphorus. The discussed modes include **empirical, analytical, semi-empirical, and artificial intelligence (AI).** The empirical mode (EM) offers simplicity and ease of operation but may struggle to meet accuracy requirements due to water quality variables' complexities and regional limitations. The analytical mode (AM) employs bio-optical models and radiation transmission models to simulate light propagation but faces challenges in practical applications. The semi-empirical mode (SEM) combines elements of empirical and analytical modes, offering better portability and retrieval accuracy. Lastly, the artificial intelligence mode (AIM) employs algorithms like neural networks and support vector machines, capturing both linear and nonlinear relationships.

Advantages:

- Hyper-spectral sensors, unmanned aerial vehicles (UAVs), and artificial intelligence contribute to advanced water quality retrieval.
- Specific retrieval algorithms for various water quality variables enhance precision.

Disadvantages:

- Empirical mode simplicity may lead to less accurate results,AI models, require extensive training data.

S.No	Article Title	Journal details	Algorithms/Models	Dataset	Advantages	Disadvantages
1	Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study	Water Research, Volume 38, Issue 18,2004,ScienceDirect	**Multivariate techniques**, including **cluster analysis (CA), factor analysis/principal component analysis (FA/PCA), and discriminant analysis (DA).**	A total of eight monitoring sites were selected in The Gomti River, a major tributary of the Ganga River system in northern India has been selected for this case study. The river originates from a natural reser	• Effective grouping of sampling sites using Cluster Analysis (CA). • Enhanced understanding of complex water quality data.	• Complex data analysis requiring statistical expertise.
2	A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques	Sensors, Volume-16, Article-1298,2016,MDPI	Evaluation and quantification of four crucial water quality parameters-chlorophyll , pH, salinity, suspended matter	-	• Diverse sensor options available on satellites and aircraft platforms. • The project investigates common approaches and sensors used for evaluating and quantifying eleven critical water quality parameters, contributing to a better understanding of water quality dynamics.	• Validation through ground measurements adds complexity and cost.
3	Assessment of Surface Water Quality Using Water Quality Index and Discriminant Analysis Method	Water,VOLUME 15,Article 680,2023,MDPI	Discriminant analysis (DA) and Water Quality Index (WQI),Irrigation Water Quality Index (IWQI)	This dataset is from the Koudiat Medouar watershed in northeastern Algeria, spanning 590 km2, and covers a two-year period (2019-2020).	• Use of internationally recognized indexes (WQI and IWQI). • Weighted parameter evaluation for precision.	• Data collection limited to a two-year period. • Oversimplification of water quality classification.

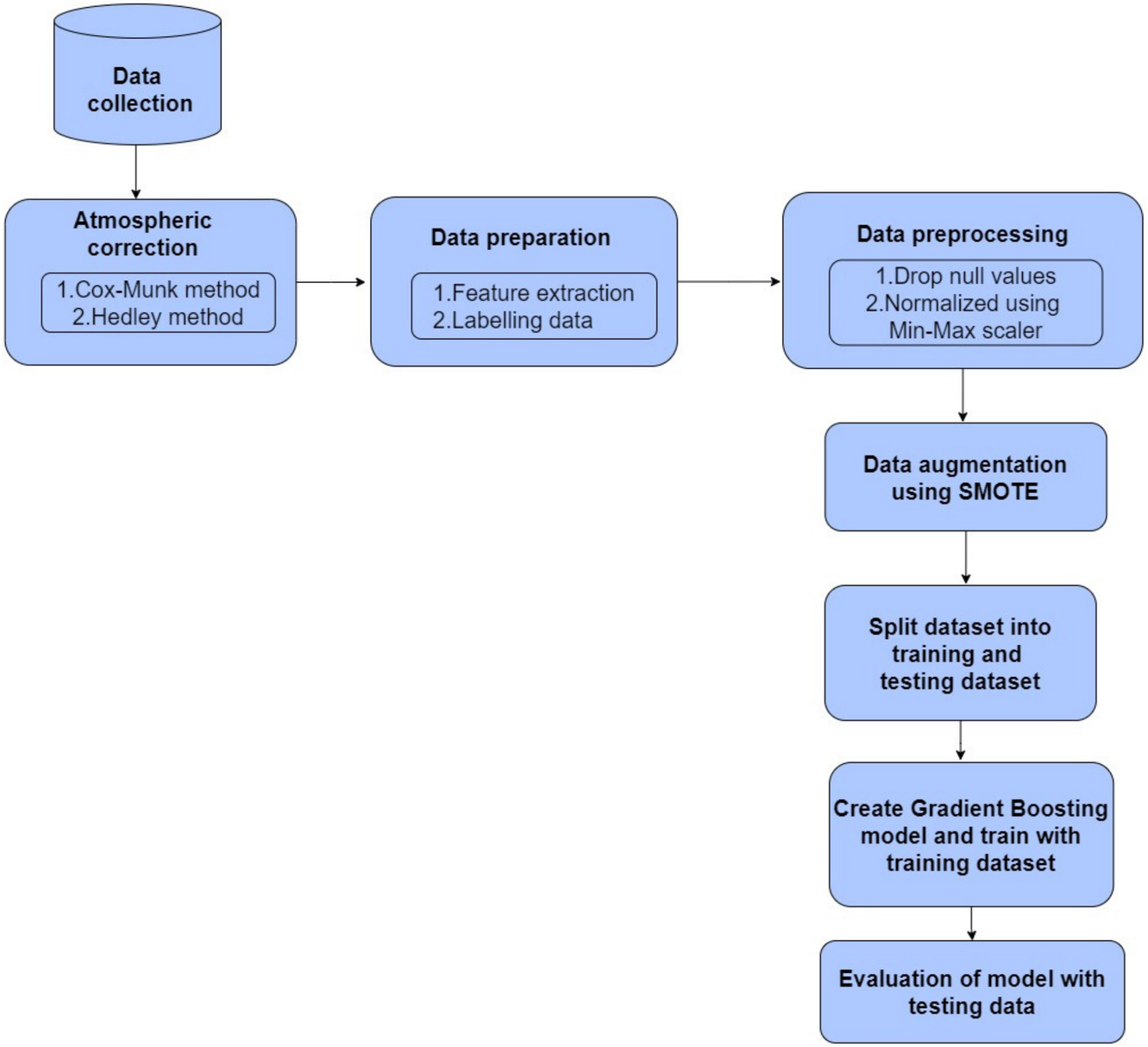
S.No	Article Title	Journal details	Algorithms/Models	Dataset	Advantages	Disadvantages
4	A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques	Chemometrics and Intelligent Laboratory Systems,volume 214,2021,ScienceDirect	Long Short Term Memory Recurrent Neural Networks (LSTM RNNs) with 91%	Tilesdit dam in Bouira, Algeria, collected over three years (2016–2018)	• Enhanced water quality classification accuracy (99.72%)	• It relies on a specific deep learning model, LSTM RNNs, which can be computationally intensive and may require significant computational resources for training and deployment.
5	Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul —Brazil	SN Applied Sciences,Volume 3,2021,SpringerLink	Multiple Linear Regression (MLR) with 85%	The Mirim Lagoon hydrographic basin in southern Brazil, comprising 154 water samples collected during 22 sampling campaigns at 7 monitoring points between 2015 and 2017.	• Cost reduction in monitoring and ease of communication with the simplified index.	• Oversimplifying water quality assessment by reducing the number of variables, which may not capture all relevant parameters accurately.
6	Remote Sensing Techniques to Assess Water Quality	Photogrammetric Engineering & Remote Sensing,volume 69,2003,ResearchGate	Empirical and physical based models	-	• Spatial and temporal monitoring capabilities. • Focus on key pollutants affecting water bodies.	• Limitations of empirical models for specific conditions. • Indirect inference for certain water quality parameters.

S.No	Article Title	Journal details	Algorithms/Models	Dataset	Advantages	Disadvantages
7	Remote sensing of water quality in an Australian tropical freshwater impoundment using matrix inversion and MERIS images	Remote Sensing of Environment,Volume 115, Issue 9,2011,ScienceDirect	Matrix Inversion Method (MIM) with a semi-analytic model with accuracy 86%	Burdekin Falls Dam in Northern Australia and measurements taken using RAMSES spectroradiometers	• By comparing different weighting schemes, the research identifies the best-performing schemes for estimating chlorophyll-a, tripton, and colored dissolved organic matter (CDOM) concentrations	• Small Sample Size
8	A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges	Remote Sensing,VOLUME 14,ARTICLE 1770,2022,MDPI	Empirical, analytical, semi-empirical, and artificial intelligence (AI)	Multispectral data like Landsat, Sentinel-2, and SPOT satellites, suitable for empirical modeling. Hyperspectral data from satellites like Hyperion and HIS.While non-satellite sources like UAV-based systems provide flexibility but may have cost and coverage limitations for large water areas.	• Hyper-spectral sensors, unmanned aerial vehicles (UAVs), and artificial intelligence contribute to advanced water quality retrieval. • Specific retrieval algorithms for various water quality variables enhance precision.	• Empirical mode simplicity may lead to less accurate results, • AI models, require extensive training data.

GAP ANALYSIS

1. Spatial Coverage Disparity
2. Lack of Atmospheric Correction
3. Inadequate Integration of Multiple Water Quality Parameters
4. Need for improved accuracy

PROPOSED MODEL



MODULES OF PROPOSED MODEL

1. Data preparation and preprocessing
2. Data augmentation
3. Training the model
4. Model evaluation
5. GUI

Module 1. Data preparation and preprocessing

Feature extraction :

- Extracting key water quality parameters that include chlorophyll, pH, dissolved oxygen, salinity, turbidity, dissolved organic matter and suspended matter from the atmospherically corrected Sentinel-2 radiance values through established research formulas is an essential step.

Labelling data :

- Labeling the extracted data involves applying threshold values to determine the water quality class, such as categorizing it as "good" or "bad" or “needs treatment”.
- This crucial step simplifies the interpretation of complex data.

Dropping null values:

- A data cleaning process that involves removing rows or entries in a dataset where certain values are missing or undefined.
- This step helps ensure the integrity and reliability of the data for analysis and prevents incomplete or inaccurate information from affecting the results.

Normalization:

- Normalization with a Min-Max scaler is a data preprocessing step that scales all values in a dataset to a specified range, often between 0 and 1.
- It ensures that features with different scales contribute equally to machine learning models, enhancing their performance and accuracy.

Module 2. Data Augmentation:

- Data augmentation using SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic examples of the minority class, effectively increasing its representation.
- So, there were balanced observations for each class.

Module 3. Model Training and Evaluation

- The next pivotal step in our project involves the creation and training of a robust machine learning model using the powerful Gradient Boosting algorithm.
- With the dataset preprocessed and features extracted, we are now ready to harness the predictive capabilities of Gradient Boosting to enhance our water quality assessment.
- Gradient Boosting, known for its ability to handle complex relationships in data, is poised to provide accurate and reliable predictions of various water quality parameters.
- Through rigorous model training and validation, we aim to harness the full potential of this algorithm to contribute significantly to our remote sensing-based water quality assessment framework.

Module 4. Designing GUI and model deployment

- This module serves as the bridge between the web application and the predictive capabilities of the model.
- Integrating a Machine Learning model into a Django application involves the thoughtful design of a dedicated module within the app.
- This design ensures that the model seamlessly fits into the application's architecture, allowing for efficient and accurate predictions while maintaining code modularity and maintainability.
- It typically encompasses functions or classes responsible for loading the pre-trained model, preprocessing user inputs, making predictions, and returning results.

ALGORITHMS/PSEUDO CODES

Algorithm 1 - Feature extraction

Step 1: Load the dataset with 13-band radiance values

Assuming you have a dataset named 'radiance_data.csv' with columns B1...B12

```
data = pd.read_csv('/content/mini2.csv')
```

Step 2: Calculate the specified features

2.1 pH calculation

```
data['pH'] = 8.339 - 0.827 * (data['B1'] / data['B8'])
```

2.2 Salinity calculation

```
data['Salinity'] = (data['B11'] - data['B12']) / (data['B11'] + data['B12'])
```

2.3 Turbidity calculation

```
data['Turbidity'] = (data['B4'] - data['B3']) / (data['B4'] + data['B3'])
```

2.4 Land Surface Temperature calculation

```
data['LS Temperature'] = data['ST_B10']*0.00341802+149.0 - 273.15
```

2.5 Chlorophyll calculation

```
data['Chlorophyll'] = (data['B5'] - data['B4']) / (data['B5'] + data['B4'])
```

2.6 Suspended Matter calculation

```
data['Suspended Matter'] = data['Oa08_radiance'] / data['Oa06_radiance']
```

2.7 Dissolved Organic Matter calculation.

```
data['Dissolved Organic Matter'] = data['Oa08_radiance'] / data['Oa04_radiance']
```

2.8 Dissolved Oxygen calculation

```
data['Dissolved Oxygen'] = -0.0167 * data['B8'] + 0.0067 * data['B9'] +
0.0083*data['B11'] + 9.577
```

Step 3: Save the dataset with extracted features to a new CSV file

```
data.to_csv('features_data.csv', index=False)
```

Algorithm 2 - Data Preprocessing

Step 1: dropping null values:

1.1 : Identify columns with missing values (NaN) i.e, count_nan = len(df) - df.count()

1.2 : Remove rows with missing values

```
data.dropna(inplace=True)
```

Step 2: Normalize the specified columns

2.1 : from sklearn.preprocessing import MinMaxScaler

2.2 : Select columns for normalization

```
columns_to_normalize = ['Dissolved Oxygen', 'Temperature', 'pH',
'Turbidity', 'Dissolved Organic Matter', 'Suspended Matter', 'Chlorophyll']
```

2.3 : Create a MinMaxScaler instance

```
scaler = MinMaxScaler()
```

2.4 : Apply scaling to the selected columns

```
scaled_data = scaler.fit_transform(data[columns_to_normalize])
```

2.5 : Create a DataFrame with scaled data

```
df_scaled = pd.DataFrame(scaled_data, columns=columns_to_normalize)
```

Step 3: Save the cleaned and normalized dataset to a new CSV file

```
df_scaled.to_csv('cleaned_normalized_data.csv', index=False)
```

Algorithm 3- labelling data - Label data based on specified criteria

Step 1 : Label data as 'good'

1.1 : good = data[
(data['Chlorophyll'] >= -0.1) & (data['Chlorophyll'] <= 0.1) &
(data['Dissolved Organic Matter'] < 500) &
(data['Dissolved Oxygen'] > 6.5) & (data['pH'] >= 6.5) & (data['pH'] <= 8.5) &
(data['Suspended Matter'] >= 300) & (data['Suspended Matter'] <= 600) &
(data['Turbidity'] >= -0.2) & (data['Turbidity'] <= 0) &
(data['Temperature'] >= 15) & (data['Temperature'] <= 35)]

Step 2: Label data as 'poor'

2.1 : poor = data[
(data['Chlorophyll'] > 0.5) |
(data['Dissolved Organic Matter'] > 2000) |
(data['Dissolved Oxygen'] < 4) |
((data['pH'] > 11) & (data['pH'] < 14)) | ((data['pH'] > 1) & (data['pH'] < 4)) |
((data['Suspended Matter'] > 900) & (data['Suspended Matter'] < 1200)) |
(data['Turbidity'] > 0.2) | (data['Temperature'] > 40) | (data['Temperature'] < 5)]

Step 3: label remaining records as ‘needs treatment’

df_all['Class'].fillna('Needs Treatment', inplace=True)

Algorithm 4 - Data Augmentation

Step 1: Import necessary libraries

```
from imblearn.over_sampling import SMOTE
```

Step 2: Initialize the SMOTE oversampler

```
oversample = SMOTE()
```

Step 3: Apply SMOTE oversampling to the training data

```
X_train_over, y_train_over= oversample.fit_resample(X_train_scaled, y_train)
```

Step 4: Apply SMOTE oversampling to the validation data (if needed)

```
X_valid_over, y_valid_over = oversample.fit_resample(X_valid_scaled,y_valid)
```

Step 5: Convert the oversampled validation arrays to DataFrames (optional)

```
X_valid_over = pd.DataFrame(X_valid_over)
y_valid_over = pd.Series(y_valid_over)
```

Algorithm 5 - Model Creation and Evalution

Step 1: Import necessary libraries

```
from imblearn.over_sampling import SMOTE
```

Step 2: Create and train the Random Forest model

```
rf = RandomForestClassifier()
rf.fit(X_train_over, y_train_over)
```

Step 3: Evaluate the model using cross-validation

```
rf_pred = cross_val_predict(rf, X_valid_over, y_valid_over, cv=5)
```

Step 4: Generate and print the classification report

```
report = classification_report(y_valid_over, rf_pred, target_names=['0', '1', '2'])
```

Algorithm 5 - Model Creation and Evalution

Step-1: Setup Django Project:

1.1: Install Django by running **pip install django.**

1.2: Create a new project by running **django-admin startproject projectname.**

1.3: Create a new app inside the project> **python manage.py startapp appname.**

Step-2: Serialize and save the trained model to a file using tools like **joblib**

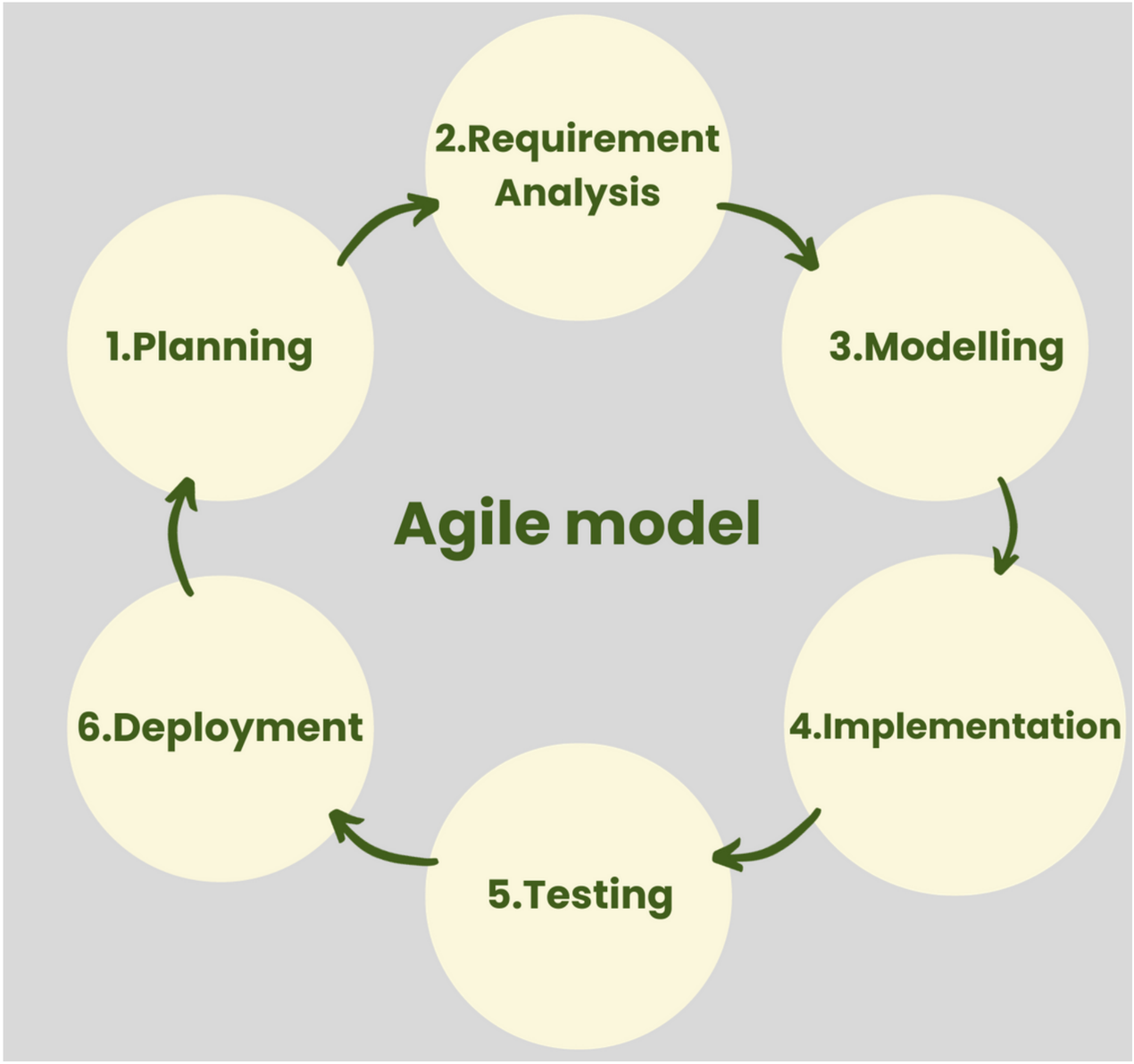
Step-3: Integrate the Model with Django

Step-4: Create Django Views and Templates

Step-5: Configure URL Patterns to route HTTP requests to the appropriate views

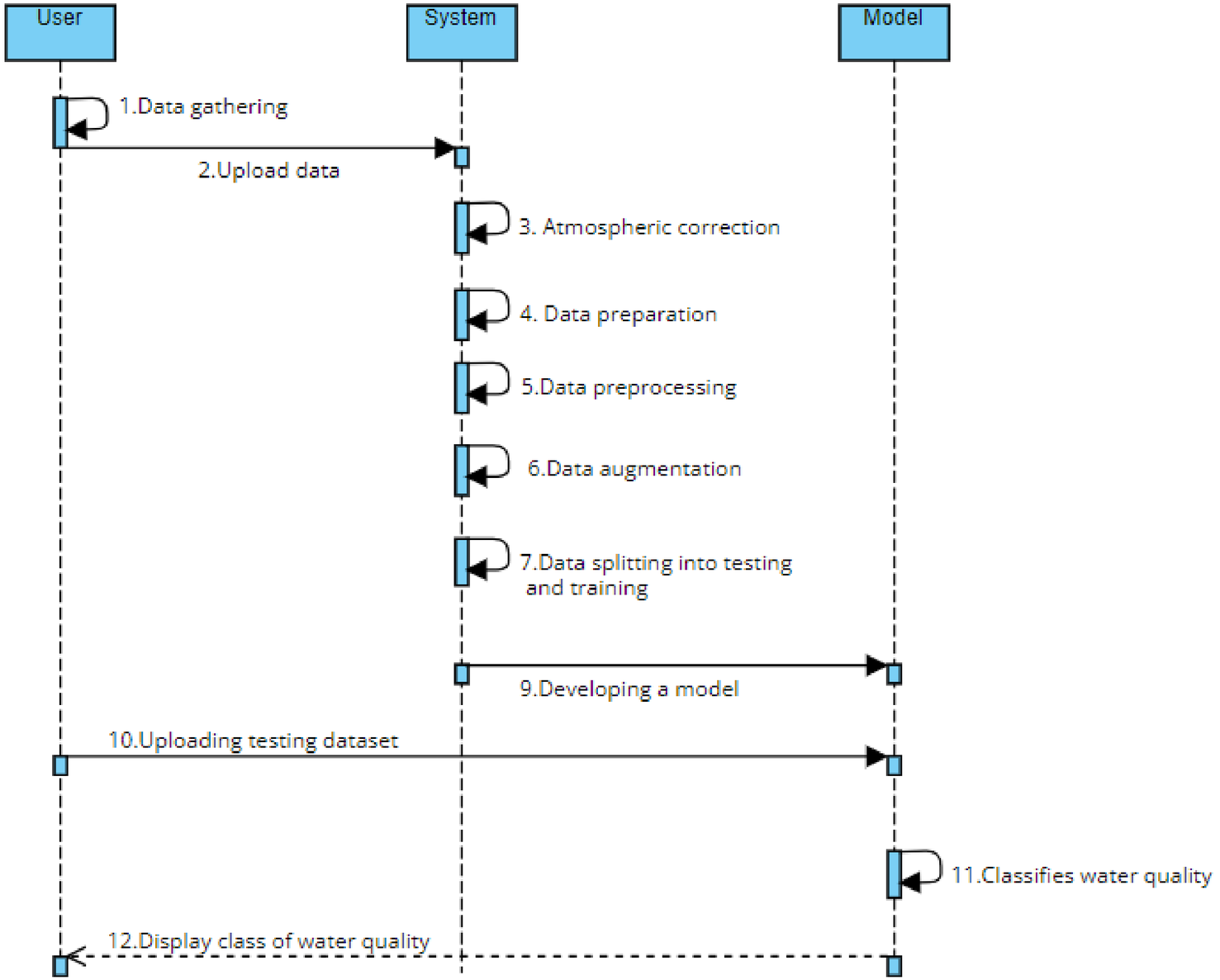
Step-6: Deploy Django application using the chosen hosting service's(eg. AWS) deployment instructions.

SDLC MODEL


2.Requirement Analysis
3.Modelling
4.Implementation
5.Testing
6.Deployment
1.Planning
Agile model

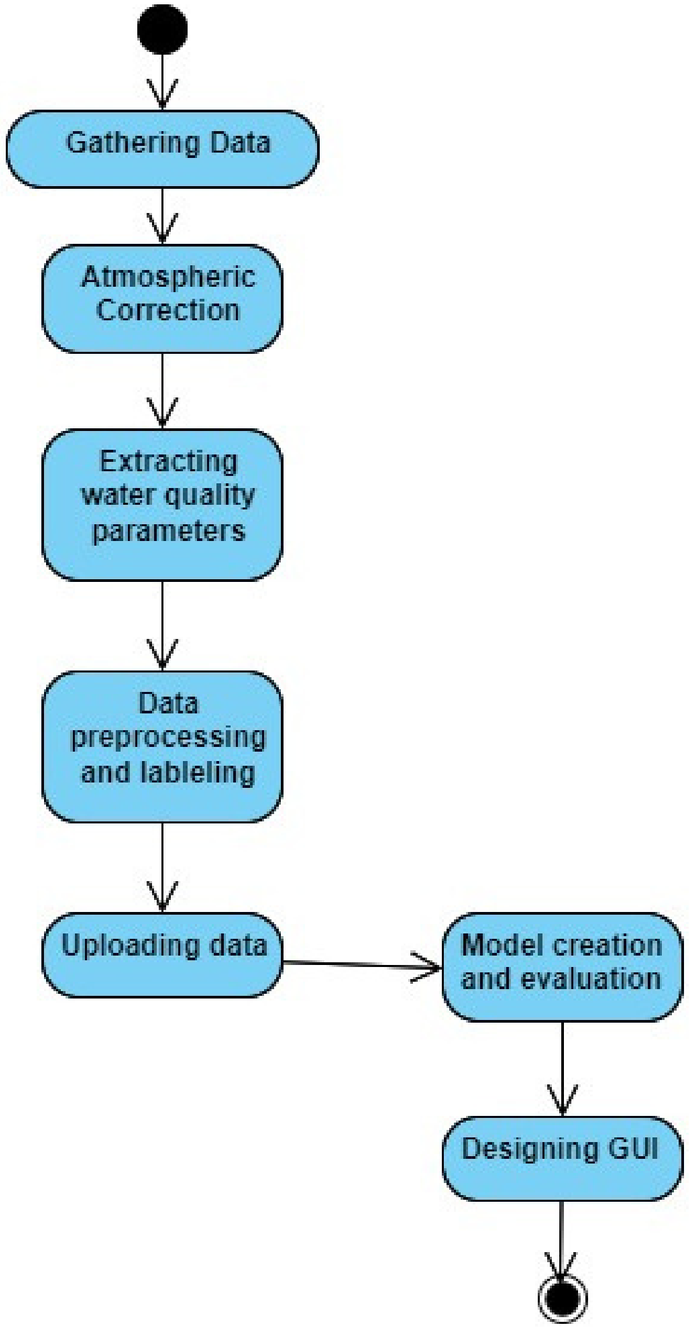
UML DIAGRAM

SEQUENCE DIAGRAM



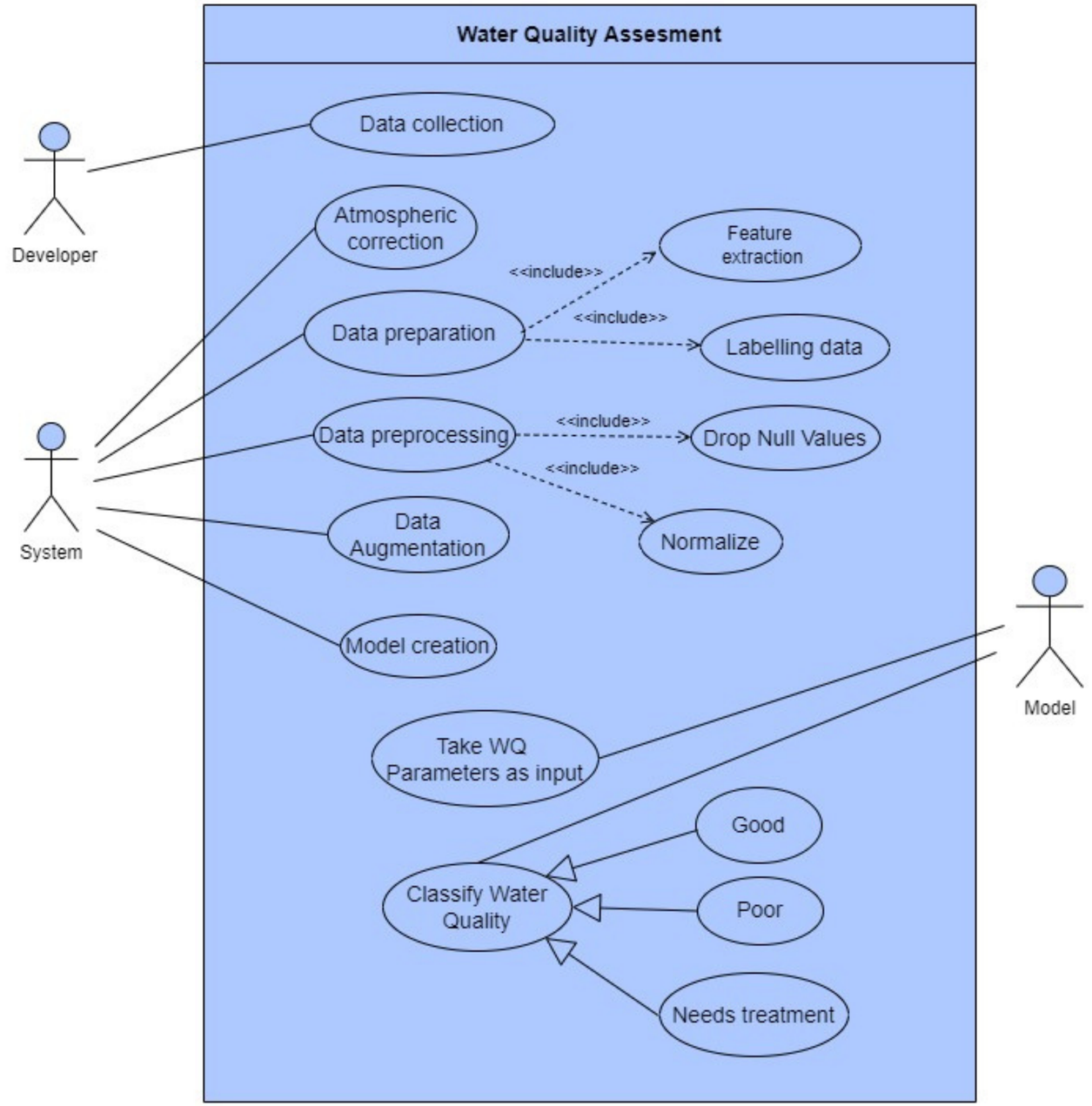
UML DIAGRAM

ACTIVITY DIAGRAM



UML DIAGRAM

USE CASE DIAGRAM



FUNCTIONAL AND NON FUNCTIONAL REQUIREMENTS

FUNCTIONAL PARAMETERS

Parameter Integration and Classification:

- The system should integrate multiple water quality parameters into a unified assessment framework.
- It should classify water quality into predefined categories (e.g., good, bad, needs treatment) based on assessment results.

NON-FUNCTIONAL PARAMETERS

Accuracy:

- The system should strive for high accuracy in predicting water quality parameters.

Scalability:

- The system should be scalable to handle large datasets and accommodate future increases in data volume.

Usability:

- The user interface should be intuitive and user-friendly, catering to both technical and non-technical users.

DATASET DESCRIPTION

Name of the Dataset: Water Quality Assessment Dataset

- **Description:** This dataset is a collection of structured information that captures various parameters and attributes that influence the quality of water in specific aquatic environments, such as rivers, lakes, or oceans.
- These datasets are crucial for environmental monitoring, research, and decision-making in areas like **water resource management, pollution control, and aquatic ecosystem health.**

Classes: 3

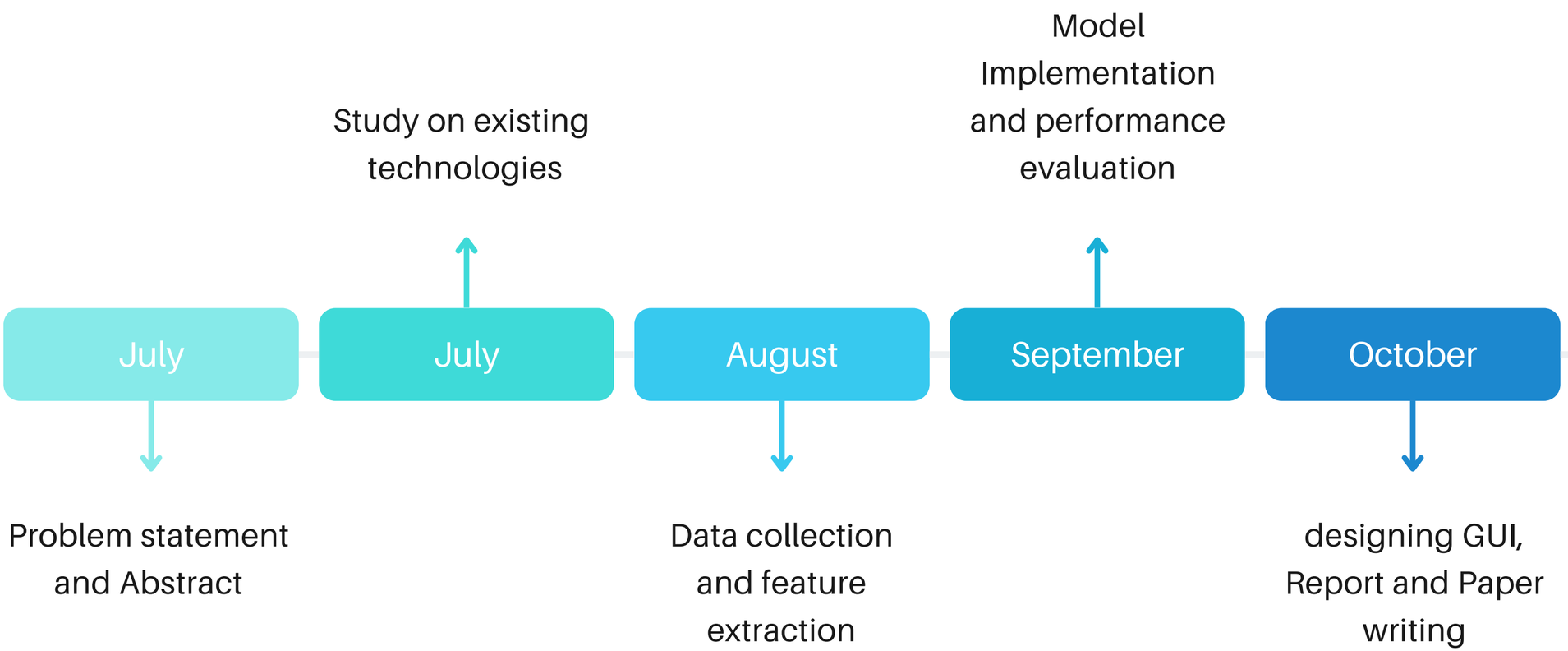
Number of records: 4200 (after upsampling)

Number of features: 8

Train set size: 80 %

Test set size: 20 %

TIMELINE CHART



REFERENCES

1. Singh, Kunwar & Malik, Amrita & Mohan, Dinesh & Sinha, Sarita. (2004). Multivariate Statistical Techniques of the Evaluation of Spatial and Temporal Variations in Water Quality of Gomti River (India) – A Case Study. Water research. 38. 3980-92. 10.1016/j.watres.2004.06.011.
2. Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. *Sensors* **2016**, *16*, 1298. https://doi.org/10.3390/s16081298
3. Mammeri, A.; Tiri, A.; Belkhiri, L.; Salhi, H.; Brella, D.; Lakouas, E.; Tahraoui, H.; Amrane, A.; Mouni, L. Assessment of Surface Water Quality Using Water Quality Index and Discriminant Analysis Method. Water 2023, 15, 680. https://doi.org/10.3390/w15040680
4. Dilmi, Smail & Ladjal, Mohamed. (2021). A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques. Chemometrics and Intelligent Laboratory Systems. 214. 104329. 10.1016/j.chemolab.2021.104329.
5. Valentini, M., dos Santos, G.B. & Muller Vieira, B. Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil. SN Appl. Sci. 3, 70 (2021). https://doi.org/10.1007/s42452-020-04005-1
6. Ritchie, Jerry & Zimba, Paul & Everitt, James. (2003). Remote Sensing Techniques to Assess Water Quality. Photogrammetric Engineering & Remote Sensing. 69. 10.14358/PERS.69.6.695.
7. Campbell, Glenn & Phinn, Stuart & Dekker, Arnold & Brando, Vittorio. (2011). Remote sensing of water quality in an Australian tropical freshwater impoundment using matrix inversion and MERIS images. Remote Sensing of Environment - REMOTE SENS ENVIRON. 115. 2402-2414. 10.1016/j.rse.2011.05.003.
8. Yang, H.; Kong, J.; Hu, H.; Du, Y.; Gao, M.; Chen, F. A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges. Remote Sens. 2022, 14, 1770. https://doi.org/10.3390/rs14081770 https://www.mdpi.com/1576850

REFERENCES

9. https://go.nasa.gov/421qeoN

10.Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. *Sensors* **2016**, *16*, 1298. https://doi.org/10.3390/s16081298

11.H. Wang, Z. Ren, M. Tang, A. Shi and F. Huang, "Design of water quality monitoring based on SVM and its simulation platform by remote sensing," 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 2010, pp. 2163-2167, doi: 10.1109/CISP.2010.5647474.

12.S. Kulkarni and V. Kelkar, "Classification of multispectral satellite images using ensemble techniques of bagging, boosting and adaboost," 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), Mumbai, India, 2014, pp. 253-258, doi: 10.1109/CSCITA.2014.6839268.


Thank You