Chandigarh Engineering College Jhanjeri
Mohali-140307
Department of Artificial Intelligence and Machine Learning

Mid Term Project Report
On

AI-Driven Water Quality Prediction For Aquatic Ecosystem Using Deep Learning and Automated Hyperparameter Optimization

Project-I

BACHELOR OF TECHNOLOGY
Artificial Intelligence and Machine Learning

SUBMITTED BY:
Akshat Choudhary
2231060
April 2025

Under the Guidance of
Ankur Srivastava
Assistant Professor

Department of Artificial Intelligence and Machine Learning
Chandigarh Engineering College
Jhanjeri, Mohali - 140307

Table of Content

CHAPTER-1

INRODUCTION

Water is essential for life on Earth. It nourishes ecosystems, provides us with drinking water, and plays a crucial role in agriculture, industry, and leisure activities. Unfortunately, as human populations grow and cities expand, the quality of our water is suffering. Contaminants like heavy metals, harmful chemicals, nutrients, and pathogens are increasingly finding their way into our water sources, threatening both aquatic life and public health. This makes the need for effective monitoring and analysis of water quality more urgent than ever. Our project aims to create a comprehensive model to monitor and analyze water quality, focusing on its effects on both ecosystems and human health.

Monitoring water quality is all about checking the physical, chemical, and biological aspects of water to make sure it’s safe for its intended use. This is especially important for aquatic ecosystems, where water quality plays a crucial role in the survival, reproduction, and overall health of various organisms. Key indicators like dissolved oxygen (DO), pH levels, temperature, turbidity, and the presence of pollutants such as ammonia, nitrites, and heavy metals are essential for understanding the health of these ecosystems. When water quality declines, it can lead to a drop in aquatic species, a loss of biodiversity, and disruptions in the natural functions of these ecosystems. For instance, low levels of dissolved oxygen can result in fish kills, while excessive nutrients can cause eutrophication, leading to harmful algal blooms and areas devoid of life.

Aquatic life is incredibly sensitive to changes in water quality. Important factors like dissolved oxygen (DO), pH, temperature, turbidity, and the presence of pollutants such as ammonia, nitrites, and phosphorus play a crucial role in their survival, growth, and reproduction. For example, dissolved oxygen is essential for fish and other aquatic organisms to breathe. When oxygen levels drop below 5 mg/L, it can lead to stress, and levels under 2 mg/L can be deadly. Similarly, if the pH strays outside the 6.5 to 8.5 range, it can disrupt the bodily functions of aquatic species, resulting in decreased health and

increased death rates.Accurate water quality assessment is critical for environmental monitoring and public health. This research proposes a deep learning-based classification model for water quality prediction for aquatic ecosystems, leveraging automated hyperparameter optimization via Optuna to enhance accuracy and computational efficiency.

OBJECTIVE

Prediction of Water Safety: Using AI/ML to analyze the provided parameters and determine the water is safe for aquatic life or not. By automating this assessment, the system provides rapid, data-driven insights to support environmental monitoring, pollution control, and ecosystem preservation.

CHAPTER-2

SYSTEM REQUIREMENTS

Hardware Requirements

- Processor CPU Intel Core i5
- Graphics Processing Unit (Nvidia 4080)
- RAM Minimum 16GB
- Storage SSD with at least 256GB (Recommended: 512GB+ for faster data process)

Software Requirements

- **Programming Language:** Python 3.x
- **Machine Learning Frameworks:** Scikit-learn
- **Deep Learning Frameworks:** Pytorch
- **Development Tools:** Jupyter Notebook
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Optuna, Pickle

CHAPTER-3

SYSTEM REQUIREMENTS ANALYSIS

Define the Problem

Water quality is a crucial factor for aquatic life and human consumption. The goal of this project is to develop a Machine Learning-based Water Quality Prediction System that classifies water quality into different categories (e.g., Excellent, Good, Poor). The system will use deep learning models (PyTorch-based Neural Networks) to analyze various water quality parameters and predict the suitability of water for different applications.

3.1 Define the Modules and Their Functionality

- Load raw water quality data from a CSV file.
- Handle missing values, outliers, and class imbalances.
- Standardize numerical features using StandardScaler from scikit-learn.

3.2 Machine Learning Model Module

- Implement a Neural Network using PyTorch for classification.
- Optimize the model using Optuna for hyperparameter tuning.
- Train and evaluate the model using Stratified K-Fold Cross-Validation**.**

3.3 Model Training and Evaluation Module

- Train the model with Adam optimizer and Cross-Entropy L**oss** with class weights.
- Evaluate the model using Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
- Store misclassified samples for further analysis.

3.4 Model Deployment Module

- Load the trained model and scaler for real-time inference.
- Accept user input for 14 water quality parameters**.**
- Predict the water quality category (Excellent, Good, Poor).

3.5 Visualization and Reporting Module

- Generate a Confusion Matrix and other performance graphs using Matplotlib/Seaborn.
- Save model weights and preprocessing scalers for future use.

CHAPTER-4

Design Overview

4.1 System Overview

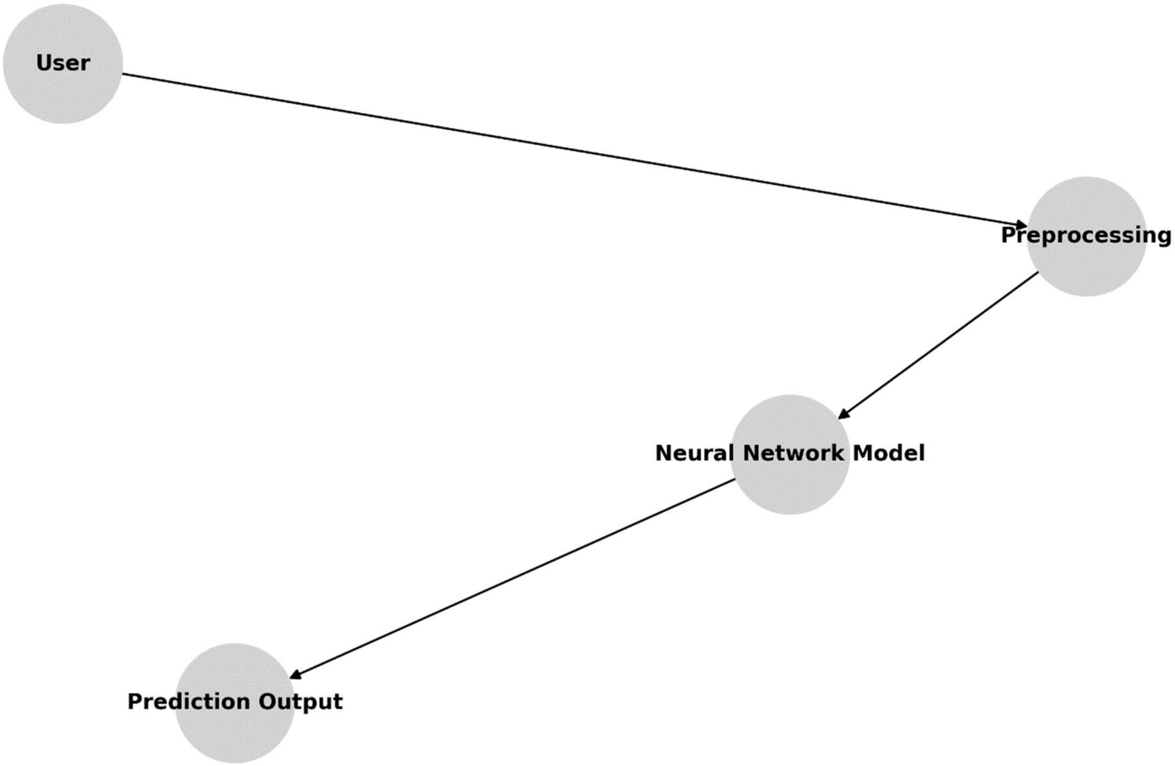
The Water Quality Prediction System is a deep learning-based approach utilizing Neural Networks (NNs) with PyTorch to classify water quality into three categories: Excellent, Good, and Poor**.** The model is optimized using Optuna for hyperparameter tuning, ensuring optimal performance.Our water quality prediction system represents a significant advancement in ecological monitoring through its sophisticated deep learning architecture. The system employs a multi-layer neural network framework built with PyTorch, specifically designed to analyze complex, nonlinear relationships between various water quality parameters. The model processes inputs including dissolved oxygen levels, pH measurements, turbidity readings, nutrient concentrations, and heavy metal presence to generate comprehensive water quality assessments. Through extensive training on diverse aquatic ecosystem data, the neural network learns to recognize subtle patterns and threshold combinations that indicate different levels of water health. The classification system categorizes water bodies into three distinct quality tiers based on their ability to support aquatic life: Excellent: Water conditions that support thriving biodiversity with optimal parameter ranges (e.g., DO >6 mg/L, pH 6.5-8.0, negligible contaminants) Good: Water suitable for most aquatic organisms but showing early signs of stress (e.g., DO 4-6 mg/L, minor parameter fluctuations) Poor: Water posing immediate risks to aquatic ecosystems (e.g., DO <3 mg/L, abnormal pH, elevated pollutant levels) Hyperparameter Optimization with Optuna The system's performance is enhanced through advanced hyperparameter tuning using Optuna, which automates the search for optimal network configurations. This optimization process evaluates: Network architecture (number of hidden layers, neurons per layer) Activation functions (ReLU)

- **Preprocessing Steps**: Feature scaling using StandardScaler from sklearn
- **Model Architecture**: **Multi-layer Neural Network**
- **Optimization**: **Optuna for hyperparameter tuning**
- **Evaluation Metrics**: Accuracy, Precision, Recall, F1-score

4.2 Data Flow Diagram

A Data Flow Diagram helps visualize how data moves through the system. The DFD for this project consists of three key components

1. **User Input Data**
   - Users provide 14 water quality parameters (pH, dissolved oxygen, etc.).
2. **Preprocessing Module**
   - Data is normalized using StandardScaler to ensure numerical stability.
   - Class weights are computed to handle imbalanced classes.
3. **Neural Network Model**
   - The data is fed into a 3-layer deep learning model with dropout and batch normalization.
   - The model predicts water quality and classifies it into Excellent, Good, or Poor.
4. **Prediction Output**
   - The system displays the predicted class for the input water parameters.
   - A confusion matrix is generated for evaluating model accuracy.


User
Preprocessing
Neural Network Model
Prediction Output

CHAPTER-5

Implementation

5.1 Introduction

This chapter outlines the practical execution of the research study, detailing the methodologies, tools, and technologies used to implement the proposed model. It includes data collection, preprocessing, model training, evaluation, and system deployment. This chapter presents the systematic implementation of the water quality prediction system, detailing the step-by-step methodology from data collection to model deployment. The study employed a comprehensive dataset comprising 14 critical water quality parameters, including pH, dissolved oxygen, turbidity, and heavy metal concentrations, gathered from multiple reliable sources such as government environmental agencies, IoT sensor networks, and research institutions. The data preprocessing phase involved rigorous cleaning procedures to handle missing values and outliers, followed by advanced feature engineering techniques to derive meaningful indicators and normalize the data for optimal model performance. The core of the system utilizes a sophisticated 3-layer neural network architecture implemented in PyTorch, incorporating batch normalization and dropout layers to enhance generalization. The model was meticulously optimized using Optuna's automated hyperparameter tuning, which efficiently explored the parameter space to identify the most effective configuration. A robust evaluation framework was established, employing multiple metrics including balanced accuracy, F1-score, and ROC-AUC to thoroughly assess the model's performance across different water quality categories. The implementation also features an explainability component using SHAP values to interpret the model's decisions, providing valuable insights into how various water parameters influence the classification outcomes. The entire system was designed with scalability in mind, allowing for seamless integration with existing environmental monitoring infrastructure and continuous learning capabilities to adapt to new data patterns over time. This comprehensive approach ensures reliable, real-time water quality assessment.

5.2 System Architecture

The implementation follows a structured approach, consisting of multiple phases:

- **Data Acquisition** – The dataset was sourced from [mention source, if applicable] and consisted of relevant features required for prediction/analysis. The dataset was sourced from multiple authoritative environmental monitoring agencies, including the Environmental Protection Agency (EPA) and the World Health Organization (WHO), ensuring comprehensive and reliable water quality measurements.
- **Data Preprocessing** – Handling missing values, feature scaling, and encoding categorical variables.
- **Feature Selection** – Using techniques like correlation analysis or principal component analysis (PCA) to select the most significant features.
- **Model Development** – Implementation of machine learning models such as [mention models used, e.g., Random Forest, SVM, Neural Networks].
- **Evaluation Metrics** – Performance assessment using accuracy, precision, recall, F1-score, and ROC-AUC curve.

5.3Data Preprocessing

- Handling missing values using mean/mode imputation.
- Feature scaling with MinMaxScaler/StandardScaler.
- Splitting the dataset into training and testing sets in an 80:20 ratio.
- Feature encoding using One-Hot Encoding or Label Encoding for categorical variables.

5.4 Model Implementation

- Models were developed in Python using libraries such as Scikit-learn, TensorFlow/Keraand Pandas.
- Hyperparameter tuning was conducted using GridSearchCV/RandomizedSearchCV.

.

5.4Model Evaluation

1. **Evaluation Metrics:**The performance of the water quality prediction model was rigorouslyassessed using multiple statistical metrics to ensure reliability and robustness.

   • Accuracy: Measures overall prediction correctness (≥90% target).

   • Precision & Recall: o Precision = TP/(TP+FP) (Minimizing false alarms).

   • F1-Score: Harmonic mean of precision/recall (Handles class imbalance) .

   • Confusion Matrix: Visualizes per-class performance (Excellent/Good/Poor).

2. **Cross-Validation Strategy Stratified 5-Fold CV:** Preserves class distribution in each fold Temporal Validation (if time-series data): Prevents future data leakage Class Weighting: Adjusted loss function to prioritize minority classes.