

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

University of Echahid Cheikh Larbi Tebessi – Tebessa
Faculty of Exact Sciences and Natural and Life Sciences
Department of Computer Science

MASTER'S THESIS

In Partial Fulfillment of the Requirements for the Degree of Master
Specialization: Systems and Multimedia

Machine Learning Approach for Water Pollution Detection

Presented by:	Lahmar Abdessalem
	Mohamed Amine Ghouari
Academic Supervisor:	Dr. Bouakkez Fatima
Academic Year:	2025-2026

Jury Members:

Prof. name	Jury President
Prof. name	Examiner
Dr. Bouakkez Fatima	Academic Supervisor

Acknowledgment

First and foremost, I praise **Allah**, the Almighty, for granting me the strength, patience, and perseverance to complete this work.

I dedicate this thesis to my beloved **parents**, whose unconditional love and endless sacrifices have been the foundation of everything I have achieved. No words can truly express my gratitude for all you have done.

To my **brothers and sisters**, thank you for your constant encouragement and for always believing in me even during the most difficult moments of this journey.

I extend my deepest appreciation to our supervisor, **Dr. Bouakkez Fatima**, for her invaluable guidance, patience, and insightful feedback throughout this project. Her expertise and dedication have been an inspiration.

Finally, I want to thank me for believing in me. I want to thank me for doing all this hard work. I want to thank me for having no days off. I want to thank me for never quitting. I want to thank me for always being a giver And tying to give more than I receive. I want to thank me for tying to do more right than wrong. I want to thank me for just being me at all times.

Résume

Cette thèse présente une approche basée sur le Machine Learning pour la détection de la pollution de l'eau à l'aide de données environnementales et physico-chimiques. L'objectif principal est de développer des modèles intelligents capables d'identifier automatiquement les risques de contamination et d'évaluer avec précision la qualité de l'eau.

L'étude s'appuie sur trois ensembles de données complémentaires : un jeu de données sur la potabilité de l'eau provenant de Kaggle, un ensemble de données multi-classes de pollution physico-chimique, ainsi qu'un benchmark industriel de séries temporelles issu du challenge GECCO 2019. Plusieurs techniques de Machine Learning ont été appliquées et comparées, notamment Random Forest, XGBoost, CatBoost, Extra Trees et des méthodes d'ensemble.

Une méthodologie complète a été adoptée, incluant le prétraitement des données, l'analyse exploratoire, l'ingénierie des caractéristiques, la gestion du déséquilibre des classes ainsi que l'évaluation des performances à l'aide de métriques telles que l'Accuracy, la Precision, le Recall, le F1-score et le ROC-AUC. Deux phases expérimentales ont été menées afin d'étudier l'impact des techniques de rééchantillonnage et des stratégies d'optimisation sur les performances des modèles.

Les résultats montrent que les approches d'ensemble basées sur les arbres de décision ainsi que les méthodes hybrides obtiennent de fortes performances pour la classification de la qualité de l'eau et la détection d'anomalies. Cette étude met également en évidence l'importance de la gestion du déséquilibre des classes et de l'ingénierie des caractéristiques dans l'amélioration de la robustesse des systèmes intelligents de surveillance.

الملخص

تتناول هذه الأطروحة مقاربة قائمة على تعلم الآلة للكشف عن تلوث المياه باستخدام البيانات البيئية والفيزيائية الكيميائية. ويتمثل الهدف الرئيسي في تطوير نماذج ذكية قادرة على تحديد مخاطر التلوث تلقائيًا وتقييم جودة المياه بدقة وفعالية.

تعتمد الدراسة على ثلاث مجموعات بيانات متكاملة: مجموعة بيانات خاصة بصلاحية مياه الشرب من منصة Kaggle، ومجموعة بيانات متعددة الفئات للتلوث الفيزيائي الكيميائي، بالإضافة إلى معيار صناعي للسلاسل الزمنية تابع لتحدي GECCO 2019. وقد تم تطبيق ومقارنة عدة تقنيات لتعلم الآلة، من بينها Forest Random وXGBoost وCatBoost وExtra Trees إضافة إلى أساليب التعلم التجميعي.

تم اعتماد منهجية متكاملة تشمل معالجة البيانات المسبقة، والتحليل الاستكشافي، وهندسة الخصائص، ومعالجة اختلال توازن الفئات، بالإضافة إلى تقييم أداء النماذج باستخدام مقاييس مثل Accuracy وPrecision وRecall وF1-score وROC-AUC.

أظهرت النتائج أن أساليب التعلم التجميعي المعتمدة على أشجار القرار، إلى جانب النماذج الهجينة، تحقق أداءً قويًا في تصنيف جودة المياه والكشف عن الحالات الشاذة. كما أبرزت الدراسة أهمية معالجة اختلال توازن الفئات وهندسة الخصائص في تحسين متانة أنظمة المراقبة الذكية.

Abstract

Water pollution remains a critical environmental and public health challenge, while traditional monitoring approaches based on manual sampling and laboratory analysis are often costly, limited in coverage, and unsuitable for real-time detection. This thesis proposes a machine learning framework for automated water pollution detection using ensemble learning, feature engineering, and adaptive class imbalance handling. Three benchmark datasets of increasing complexity are investigated: binary water safety classification, multi-class pollution severity prediction, and Industrial IoT time-series anomaly detection.

The proposed experimental pipeline evaluates models under both raw and resampling-augmented training conditions, combining imbalance mitigation techniques with Bayesian hyperparameter optimization. Experimental results show that ensemble models, particularly gradient-boosting methods, consistently achieve strong predictive performance across all datasets and outperform the benchmark deep learning approach used for comparison.

An important finding of this study is that synthetic oversampling techniques may reduce performance when applied to engineered feature spaces or models with built-in cost-sensitive learning capabilities. Overall, the results demonstrate that optimized ensemble methods combined with principled feature engineering provide a robust, interpretable, and scalable solution for intelligent water quality monitoring systems.

Keywords: Water pollution detection, machine learning, ensemble methods, class imbalance, anomaly detection, water quality classification, feature engineering, SMOTE, gradient boosting, IoT sensor data.

Contents

List of Figures ix

List of Tables x

List of Abbreviations xi

General Introduction 1

1 Theoretical Foundations: Water Quality and the Monitoring Problem 3
- 1.1 Introduction . . . 3
- 1.2 The Hydrological System and Pollution Pathways . . . 3
  - 1.2.1 The Hydrological Cycle . . . 3
  - 1.2.2 Freshwater Sources and Their Monitoring Implications . . . 4
  - 1.2.3 Contamination Pathways and Propagation Dynamics . . . 5
- 1.3 Water Quality Characterization . . . 5
- 1.4 Standards, Thresholds, and the Water Quality Index . . . 6
  - 1.4.1 WHO Guidelines for Drinking-water Quality . . . 7
  - 1.4.2 Regional and National Regulatory Frameworks . . . 7
  - 1.4.3 The Water Quality Index (WQI) Concept . . . 7
- 1.5 Traditional vs. Automated Water Monitoring . . . 8
  - 1.5.1 Limitations of Manual Sampling . . . 8
  - 1.5.2 The Case for Automation . . . 9
- 1.6 Machine Learning: Definitions and Methods . . . 9
  - 1.6.1 Definition and Core Concepts . . . 9
  - 1.6.2 Learning Paradigms . . . 9
  - 1.6.3 ML Methods Relevant to Water Quality . . . 10
  - 1.6.4 Evaluation Metrics . . . 11
- 1.7 Conclusion . . . 12

2 Related Work 14
- 2.1 Introduction . . . 14
- 2.2 Related Works . . . 14
  - 2.2.1 Dataset 1 – Kaggle Water Quality / Potability Tabular Datasets . . . 14
  - 2.2.2 Dataset 2 – Groundwater Quality and Potential Datasets . . . 17
  - 2.2.3 Dataset 3 – GECCO 2019 Industrial Challenge Drinking-Water Time Series . . . 20
- 2.3 Comparative Synthesis . . . 22
- 2.4 Comparative Synthesis . . . 24
  - 2.4.1 Dataset 1 - Tabular Water Potability Classification . . . 24

2.4.2 Dataset 2 - Groundwater Quality and Potential Prediction 24
2.4.3 Dataset 3 - Time-Series Anomaly Detection 25
2.4.4 Cross-Dataset Methodological Trends 25
2.4.5 Persistent Research Gaps 26
2.4.6 Positioning of the Present Thesis 26
2.5 Conclusion 26

3 Methodology and Data Exploration **27**
3.1 General Methodology Overview 27
3.1.1 Experimental Pipeline: From Preprocessing to Evaluation 27
3.1.2 Feature Exploration Methods 28
3.1.3 Methodological Framework and Pipeline Architecture 29
3.1.4 Feature Engineering Methods 33
3.2 Dataset 1: Urban Water Quality (Kaggle) 35
3.2.1 Dataset Description 35
3.2.2 Exploratory Data Analysis (EDA) 36
3.3 Dataset 2: Physicochemical Pollution 40
3.3.1 Dataset Description 40
3.3.2 Exploratory Data Analysis (EDA) 41
3.4 Dataset 3: GECCO Industrial IoT Time-Series 45
3.4.1 Dataset Description 46
3.4.2 Exploratory Data Analysis (EDA) 46

4 Results and Discussion **50**
4.1 Introduction 50
4.2 Dataset 1: Binary Water Safety Classification 50
4.2.1 Dataset Characteristics and the Challenge of Class Imbalance . . . 50
4.2.2 Feature Engineering: From 20 Raw Features to 40 Discriminative Signals 51
4.2.3 Phase 1 Results: Training Without Imbalance Mitigation 51
4.2.4 Phase 2 Results: Training With Resampling 52
4.3 Dataset 2: Multi-Class Pollution Level Classification 55
4.3.1 Dataset Characteristics and the Extreme Imbalance Challenge . . . 55
4.3.2 Feature Engineering: Constructing Composite Pollution Indices . . 56
4.3.3 Phase 1 Results: With Imbalance Handling (SMOTE-ENN) 56
4.3.4 Phase 2 Results: Without Resampling (Cost-Sensitive Native Handling) 59
4.4 Dataset 3: Real-Time Anomaly Detection in Drinking Water (GECCO IIoT) 62
4.4.1 Dataset Characteristics: Large-Scale Temporal Imbalance 62
4.4.2 Feature Engineering: Temporal Signal Characterization 63
4.4.3 Phase 1 Results: Strong Baseline Without Imbalance Mitigation . . 63
4.4.4 Phase 2 Results: Imbalance Mitigation and Optuna Tuning 65
4.5 Comparison with State-of-the-Art Literature 67
4.5.1 Overview of the Comparison Framework 67
4.5.2 Dataset 1: Comparison Against Patel et al. (2022) and Fatima et al. 68
4.5.3 Dataset 2: Comparison Against Chen et al. (2024) and Elmotawakkil et al. (2025) 68
4.5.4 Dataset 3: Comparison Against Wang et al. (2025) 69

4.6 Conclusion 69

General Conclusion **71**

List of Figures

1.1 The Hydrologic Cycle. . . . 4
1.2 Manual water sampling. . . . 8
1.3 Wireless sensor capture device. . . . 8
1.4 Standard supervised learning pipeline. . . . 9
1.5 Standard binary confusion matrix 11
1.6 Receiver Operating Characteristic (ROC) curve 12

3.1 Workflow Pipeline Diagram 28
3.2 Dataset 1 Water Quality Distribution per Target 36
3.3 Dataset 1 Histograms & KDE Plots 37
3.4 Dataset 1 Correlation Heatmap 38
3.5 Dataset 1 Boxplots by Event 39
3.6 Dataset 1 Pair Plots 40
3.7 Dataset 2 Water Quality Distribution per Target 42
3.8 Dataset 2 Histograms & KDE Plots 43
3.9 Dataset 2 Correlation Heatmap 44
3.10 Dataset 2 Boxplots by Event 44
3.11 Dataset 2 Pair Plots 45
3.12 Dataset 3 Water Quality Distribution per Target 47
3.13 Dataset 3 Histograms & KDE Plots 47
3.14 Dataset 3 Correlation Heatmap 48
3.15 Dataset 3 Boxplots by Event 48
3.16 Dataset 3 Pair Plots 49

4.1 Phase 1 Model Comparison Plot 52
4.2 Phase 1: Confusion Matrix, ROC Curve, and Precision-Recall Curve 52
4.3 Class Distribution Before vs. After SMOTETomek 53
4.4 Phase 2 Model Comparison Plot 54
4.5 Phase 2: Confusion Matrix, ROC Curve, and Precision-Recall Curve 54
4.6 Best Model Metrics Comparison Between Phase 1 and Phase 2 55
4.7 Dataset 2, Phase 1: Model Performance Summary Chart 57
4.8 Test Set Results: CatBoost Model in Phase 1 58
4.9 Test Set Results: Stacking Ensemble in Phase 1 58
4.10 Dataset 2, Phase 2: Model Performance Summary Chart 59
4.11 Test Set Results: CatBoost Model in Phase 2 60
4.12 Test Set Results: Stacking Ensemble in Phase 2 61
4.13 F1 Macro and Balanced Accuracy Comparison Between Phase 1 and Phase 2 62
4.14 Phase 1: ROC and Precision-Recall Curves 64
4.15 Phase 1: Comparison of Each Metric Across All Models 64

4.16 Phase 1: Confusion Matrix for Best Model (Extra Trees) on Test Set . . . 65
4.17 Phase 2: ROC and Precision-Recall Curves 66
4.18 Phase 2: Comparison of Each Metric Across All Models 66
4.19 Phase 2: Confusion Matrix for Best Model on Test Set 67

List of Tables

1.1 Comparative characteristics of freshwater source types relevant to monitoring system design. 4
1.2 Water pollution categories, proxy parameters, and relevant machine learning data modalities. 6
1.3 Comparison of common Machine Learning methods used in Water Quality Detection. 10

2.1 Comparative summary of the five related works grouped by dataset (D1, D2, D3). 23

3.1 Urban Water Quality Parameters and Hazard Thresholds (adapted from [36]) 36
3.2 Physicochemical Pollution Dataset Parameters 41
3.3 GECCO IIoT Sensor Parameters 46

4.1 Dataset 1, Phase 1: Individual Model Performance Comparison 52
4.2 Dataset 1, Phase 2: Individual Model Performance Comparison 53
4.3 Dataset 1: Phase 1 vs. Phase 2 Final Champion Comparison 55
4.4 Dataset 2, Phase 1: Model Performance Summary (SMOTE-ENN) 57
4.5 Dataset 2, Phase 2: Model Performance Summary (No Resampling) 59
4.6 Dataset 2: Phase 1 vs. Phase 2 — CatBoost Test-Set Performance 61
4.7 Dataset 3, Phase 1: Model Performance on Raw Imbalanced Data 63
4.8 Dataset 3, Phase 2: Model Performance with BorderlineSMOTE 65
4.9 Dataset 3: Final Cross-Configuration Comparison (Selected Models) 67

List of Abbreviations

Abbreviation		Full Term
AI	—	Artificial Intelligence
ANN	—	Artificial Neural Network
BOD	—	Biochemical Oxygen Demand
CNN	—	Convolutional Neural Network
COD	—	Chemical Oxygen Demand
CSV	—	Comma-Separated Values
DL	—	Deep Learning
DO	—	Dissolved Oxygen
DT	—	Decision Tree
EC	—	Electrical Conductivity
GPU	—	Graphics Processing Unit
IoT	—	Internet of Things
KNN	—	K-Nearest Neighbours
LSTM	—	Long Short-Term Memory
MAE	—	Mean Absolute Error
ML	—	Machine Learning
MSE	—	Mean Squared Error
NB	—	Naïve Bayes
NTU	—	Nephelometric Turbidity Unit
pH	—	Potential of Hydrogen
RF	—	Random Forest
RMSE	—	Root Mean Squared Error
RNN	—	Recurrent Neural Network
ROC	—	Receiver Operating Characteristic
SVM	—	Support Vector Machine
TDS	—	Total Dissolved Solids
TH	—	Total Hardness
TSS	—	Total Suspended Solids
WHO	—	World Health Organization
WQI	—	Water Quality Index
WTP	—	Water Treatment Plant
XGB	—	eXtreme Gradient Boosting
SMOTE	—	Synthetic Minority Over-sampling Technique

General Introduction

Water is one of the most essential resources for human life, yet its quality is increasingly threatened by industrial discharge, agricultural runoff, urbanization, and climate change. According to the World Health Organization, millions of people worldwide still lack access to safe drinking water, highlighting the urgent need for efficient and intelligent water quality monitoring systems.

1. Motivation and Objectives

Traditional monitoring methods rely mainly on manual sampling and laboratory analysis, which are accurate but expensive, time-consuming, and unable to provide continuous real-time monitoring. These limitations motivate the development of automated systems based on Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) techniques. The main objective of this thesis is to develop and evaluate a machine learning framework for automated water pollution detection and water quality prediction. This work focuses on comparing several intelligent models, studying the impact of class imbalance, applying feature engineering techniques, and evaluating advanced ensemble methods such as Random Forest, XGBoost, CatBoost, LightGBM, and Extra Trees.

2. Problem Statement and Challenges

Applying machine learning to water pollution detection involves several important challenges. First, water quality datasets are highly imbalanced, where polluted samples are much rarer than normal ones, making accurate anomaly detection difficult. Second, raw physicochemical parameters alone are often insufficient to represent the complex interactions between contaminants.

Another challenge concerns multi-class pollution severity prediction, where distinguishing between low, medium, and high contamination levels becomes more difficult due to severe imbalance. In addition, contamination events evolve over time, requiring models capable of capturing temporal dynamics and trends. Finally, synthetic oversampling techniques such as SMOTE may sometimes generate unrealistic samples that negatively affect model performance.

3. Contributions

This thesis makes several important contributions to intelligent environmental monitoring:

1. Development of a dual-phase experimental framework combining raw and imbalance-mitigated training strategies.
2. Implementation and comparison of multiple machine learning and ensemble models for water quality prediction.
3. Achievement of high classification performance on three benchmark datasets.
4. Demonstration that oversampling techniques such as SMOTE may degrade performance in certain cases when combined with advanced ensemble methods.
5. Proposal of a domain-driven feature engineering methodology using interaction ratios, temporal statistics, and composite pollution indices.
6. Comparative evaluation of fourteen machine learning models, identifying **CatBoost** and **Extra Trees** as highly effective approaches for imbalanced water quality datasets.

4 Impact and Future Perspectives

The proposed framework contributes to improving intelligent water monitoring systems by enabling accurate and real-time pollution detection. From a scientific perspective, this work demonstrates that optimized ensemble methods combined with feature engineering can achieve state-of-the-art performance without relying on highly complex deep learning architectures.

From a practical perspective, the proposed models are computationally efficient, scalable, and suitable for real-world deployment in smart environmental monitoring systems. The obtained results may support public health protection, environmental sustainability, and automated early-warning systems for water contamination.

Future research directions include IoT integration, real-time streaming analysis, uncertainty quantification, semi-supervised learning, edge computing deployment, and geospatial environmental monitoring.

5. Organisation of the Thesis

This thesis is organised into four chapters:

Chapter 1 Presents the theoretical foundations of water quality monitoring, pollution sources, hydrological systems, and machine learning concepts.

Chapter 2 Reviews related works and recent studies concerning intelligent water quality prediction and pollution detection.

Chapter 3 Describes the proposed methodology, data preprocessing procedures, feature engineering techniques, and the experimental framework.

Chapter 4 Presents the experimental results, comparative analysis, discussion of findings, limitations, and future research perspectives. Finally, the thesis concludes with a summary of the major contributions and outcomes of the study.

Chapter 1

Theoretical Foundations: Water Quality and the Monitoring Problem

1.1 Introduction

Water is a vital resource for human existence; however, it can serve as a conduit for numerous waterborne diseases if not adequately treated or monitored. Its integrity is perpetually threatened by diverse pollution sources, including agricultural runoff, industrial discharge, and the escalating impacts of climate change. Consequently, to preserve this invaluable resource, robust and effective water quality monitoring systems must be established. This chapter commences with an exploration of the concept of water quality, the primary sources of contamination, and their documented effects. It subsequently provides a summary of current monitoring methodologies, encompassing the measurement principles and communication technologies that underpin them.

1.2 The Hydrological System and Pollution Pathways

1.2.1 The Hydrological Cycle

The **hydrological cycle** is the continuous trajectory of water as it undergoes phase changes and moves through the Earth's atmosphere, surface, and subsurface [34]. This interconnected system operates through five primary stages: **evaporation** (phase transition of liquid water to vapor driven by thermal energy), **condensation** (conversion of vapor back to liquid droplets, forming clouds), **precipitation** (delivery of water back to the surface as rain or snow), **infiltration** (percolation of water into the soil and underlying geological layers), and **surface runoff** (lateral flow of unabsorbed water over land surfaces).

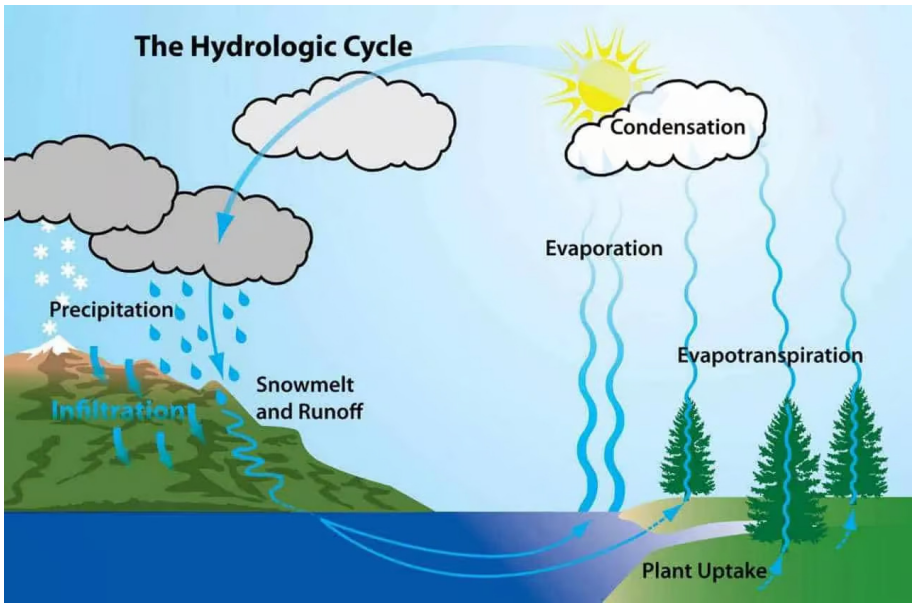


Figure 1.1: The Hydrologic Cycle.

Each stage introduces a distinct **contamination pathway**. Atmospheric pollutants from industrial emissions are deposited into surface waters via **wet and dry deposition** during precipitation [26]. Surface runoff transports dissolved **chemicals** and **suspended solids** from land into rivers and lakes. Infiltration carries soluble **contaminants** into the subsurface, potentially reaching **groundwater reserves** [26]. Because contamination can enter the water system at any stage and propagate dynamically, effective **monitoring** requires spatially distributed, continuous **data collection** rather than a single fixed measurement point.

1.2.2 Freshwater Sources and Their Monitoring Implications

Freshwater is accessed through three structurally distinct source types, summarized in Table 1.1. Each differs in replenishment dynamics, contamination behavior, and the monitoring modality it admits. Only approximately 2.5% of Earth's total water volume is fresh, with the majority locked in glaciers and deep aquifers [7].

Source Type	**Replenishment Dynamics**	**Contamination Behaviour**	**Primary Monitoring Modality**
Surface water	Rapid; precipitation and runoff	Pulse events to gradual accumulation	In-situ sensors; satellite remote sensing
Groundwater	Slow; infiltration over years to centuries	Gradual; often latent until extraction	In-situ chemical sampling at wells
Atmospheric water	Transient; vapor or droplets	Vector via wet/dry deposition	Precipitation chemistry; air quality networks

Table 1.1: Comparative characteristics of freshwater source types relevant to monitoring system design.

1.2.3 Contamination Pathways and Propagation Dynamics

Each stage of the hydrological cycle introduces a distinct **contamination pathway**:

- **Precipitation pathway** — atmospheric pollutants from industrial emissions are deposited into surface waters via wet and dry deposition [26].
- **Surface runoff pathway** — runoff transports dissolved chemicals and **suspended solids** from land into rivers and lakes.
- **Infiltration pathway** — soluble **contaminants** migrate into the subsurface, potentially reaching groundwater reserves [26].

Industrial effluents and mine drainage represent dominant **point sources** of **heavy metal contamination**, while diffuse agricultural runoff introduces **pesticide residues**, nutrients, and naturally leached metals into surface and groundwater systems [5].

The temporal dynamics of these pathways vary considerably: point-source discharges produce sharp **concentration spikes** detectable at high **sampling frequency**, whereas diffuse contamination evolves over months or years and requires spatially extensive observation. This heterogeneity positions **pollution detection** as a fundamentally **multi-scale, multi-modal data** problem — one that motivates the machine learning approaches developed in subsequent chapters.

1.3 Water Quality Characterization

Potable water is defined as water that is safe for human consumption, presenting no significant health risks over a lifetime of use. Globally, the stakes of water quality are immense: as of 2022, an estimated 2.2 billion people still lacked access to safely managed drinking water services [25].

To evaluate potability, specific physicochemical parameters serve as measurable signals. These parameters form the input feature vector for the machine learning models developed in this study. The primary parameters include:

- **pH**: A logarithmic measure of acidity or alkalinity, conventionally presented on a 0–14 scale for most natural and treated waters, though values can fall outside this range under extreme conditions such as acid-mine drainage. Anomalous pH values serve as indicators of Chemical contamination or industrial discharge.
- **Turbidity**: A measure of light scattering caused by suspended particles in the water column. High turbidity signals the presence of sediment, potential microbial contamination, or water treatment failure.
- **Dissolved Oxygen (DO)**: The concentration of free oxygen dissolved in the water. Low DO values indicate the decomposition of organic matter, thermal stratification, or algal bloom cycling, all of which indicate severe ecosystem stress.
- **Biological Oxygen Demand ($\mathbf{BOD_5}$)**: The amount of oxygen consumed by micro-organisms over a five-day incubation period at 20℃ while decomposing organic matter. A high BOD_5 value indicates a heavy organic pollution load.

- **Electrical Conductivity**: The ability of water to conduct an electrical current, serving as a proxy for total dissolved ionic solids. Elevated conductivity indicates the influx of dissolved salts, metals, or industrial effluents.
- **Heavy Metals**: Concentrations of elements such as lead (Pb), arsenic (As), mercury (Hg), and cadmium (Cd). These substances are toxic at trace concentrations, persist in biological systems through bio-accumulation, and in several cases have no established safe threshold [26].
- **Nitrates (NO_3^-) and Phosphates (PO_4^{3-})**: Dissolved nutrient ion concentrations. Elevated levels primarily signal agricultural runoff, contributing to eutrophication. In freshwater systems, phosphorus is the primary limiting nutrient and therefore the dominant driver of eutrophication, while nitrate contributes as a secondary nutrient and additionally poses a direct human health risk through methemoglobinemia at elevated concentrations [26, 9, 37].

The parameters above can be grouped into broader pollution categories that differ in their sources, temporal dynamics, and measurable signatures. Table 1.2 summarizes this taxonomy together with the corresponding machine learning data modalities. No single sensor or method covers all categories — the heterogeneity motivates the multi-modal approach pursued in this thesis.

Category	**Key Pollutants / Agents**	**Proxy Parameter(s)**	**ML Data Modality**
Chemical	Heavy metals (Pb, As, Hg), pesticides, industrial solvents	Conductivity, pH	Tabular sensor data
Biological	Pathogenic bacteria (*E. coli*), viruses, protozoa	BOD, turbidity, DO	Time-series sensor data
Physical	Suspended sediment, microplastics, thermal discharge	Turbidity; optical / microscopy imagery	Tabular sensor data; CNN image classification
Nutrient	Nitrogen and phosphorus compounds	Nitrate, phosphate, chlorophyll-*a*	Tabular sensor data; satellite remote sensing
Emerging	PPCPs, PFAS, EDCs	Spectral anomalies, trace chemical signatures	Spectroscopic analysis; anomaly detection

Table 1.2: Water pollution categories, proxy parameters, and relevant machine learning data modalities.

1.4 Standards, Thresholds, and the Water Quality Index

The parameters introduced in Section 1.3 acquire regulatory and computational meaning only once they are paired with quantitative thresholds. This section describes the

international and regional frameworks that define those thresholds, introduces the Water Quality Index as an aggregate scoring approach, and reformulates these constructs as the supervisory signal required for machine learning.

1.4.1 WHO Guidelines for Drinking-water Quality

The international benchmark for drinking-water quality is the World Health Organization (WHO) *Guidelines for Drinking-water Quality*, which establishes maximum permissible concentrations for contaminants based on epidemiological and toxicological evidence [36]. Representative limits include a nitrate ceiling of 50 mg/L, an arsenic limit of 10 µg/L, and an acceptable pH range of 6.5 to 8.5 [36]. These guidelines serve as the global reference standard against which national and regional regulations are typically calibrated.

1.4.2 Regional and National Regulatory Frameworks

At the regional level, frameworks such as the European Water Framework Directive provide statutory limits that are periodically reviewed and may be stricter or more permissive than WHO benchmarks depending on local context [17, 10]. National regulators frequently adopt WHO values as a baseline and adjust specific limits according to regional water chemistry, public health priorities, and treatment infrastructure. Despite this jurisdictional variation, the structure of these frameworks is consistent: each defines a set of parameters and an associated permissible range or maximum concentration, producing a binary or multi-class compliance status for any given sample.

1.4.3 The Water Quality Index (WQI) Concept

The **Water Quality Index (WQI)** is a **non-dimensional aggregate metric** that condenses a set of physicochemical parameters into a single composite score, supporting a categorical assessment of water quality [8]. Its general formulation combines **normalised sub-indices**, each corresponding to a single parameter scaled against its regulatory threshold, into a weighted aggregate:

$$\mathrm{WQI} = \sum_{i=1}^{n} W_i \cdot SI_i, \tag{1.1}$$

where W_i is the weight of parameter i and SI_i is its sub-index value, derived from a parameter-specific quality curve [8]. **No single universal formulation exists**: jurisdictionally specific variants such as the National Sanitation Foundation WQI (NSFWQI), the Canadian Water Quality Index (CWQI), and the British Columbia Water Quality Index (BCWQI) differ in their parameter sets, weighting schemes, and aggregation rules, and the choice of variant materially affects the resulting score [8]. The construction's principal appeal is **interpretability**: a single scalar value is readily understood by non-specialist stakeholders, and its quality categories map directly onto management decisions.

Manual WQI computation is **labour-intensive and error-prone**, as sub-index calculations are lengthy, inconsistent across competing equations, and propagate measurement uncertainty from the input parameters [8, 1]. This has motivated **machine learning models that bypass the explicit sub-index calculation**, producing a WQI estimate directly from raw parameter inputs [8, 1]. The WQI is methodologically significant in three respects:

- It provides a **continuous regression target** derived from physically meaningful inputs, enabling supervised learning even where binary potability labels are unavailable [8].
- Its categorical version provides a **multi-class classification target** capturing intermediate degradation states between *safe* and *contaminated*; tree-based ensembles have been shown to reach high accuracy on this task [1].
- Benchmarking model predictions against the WQI distinguishes models that reproduce parameter-threshold logic from those that learn richer non-linear interactions absent from the index formula.

1.5 Traditional vs. Automated Water Monitoring

Regulatory thresholds and labeling schemes only become useful once data has been measured. The practical question of **how, where, and how often** water quality is observed therefore conditions everything a downstream machine learning system can do. This section briefly contrasts manual sampling with automated monitoring to motivate the data regime assumed throughout this thesis.



Figure 1.2: Manual water sampling.



Figure 1.3: Wireless sensor capture device.

1.5.1 Limitations of Manual Sampling

Conventional water quality assessment relies on grab-sampling followed by laboratory analysis — titration, spectrophotometry, atomic absorption spectroscopy — which remains analytically precise and is retained as a regulatory reference [4]. As a primary data source for an automated detection pipeline, however, manual sampling is unsuitable on three grounds. It is **labor-intensive and costly**, which forces sparse sampling intervals that miss event-driven contamination passing between visits [23]. It is **spatially limited**, since coverage is constrained by available personnel and accessibility. And it is **prone to human error**: aggregate metrics such as the Water Quality Index require sub-index calculations across heterogeneous parameters, a manual process whose error sensitivity has been documented in operational settings [1]. The data such pipelines produce is therefore too sparse, too unevenly distributed, and too vulnerable to operator variability to support large-scale pattern-recognition tasks.

1.5.2 The Case for Automation

Automated monitoring addresses each of these limitations directly. Networked sensors generate **continuous, high-frequency** measurements without recurring human cost, capturing transient contamination events that manual sampling cannot. Distributed deployments provide **spatial coverage** that scales with hardware rather than personnel. And because measurements are recorded digitally and consistently, the operator-driven error introduced by manual sub-index computation is removed from the pipeline [24].

The resulting data regime — continuous, multi-dimensional, and digitally captured — is precisely the input format for which supervised classification, time-series modeling, and ensemble methods are designed. The shift from manual to automated monitoring is therefore not merely an engineering improvement; it is the precondition under which a machine-learning approach to water quality detection becomes feasible at all. The remainder of this thesis is concerned with the methods that exploit it.

1.6 Machine Learning: Definitions and Methods

1.6.1 Definition and Core Concepts

Machine learning (ML) is a key part of artificial intelligence. It uses algorithms to look at data and find patterns. This helps predict new information [38]. Unlike traditional programming, which uses set rules and formulas, ML uses data to solve complex problems, especially in environmental systems [38]. The process starts with training datasets. These datasets have examples or samples. Each sample has features (like sensor readings) and often labels (like water quality classification) [13, 38]. Algorithms learn from these datasets during training. Then, they are tested to see how well they can understand new, unseen data [13].

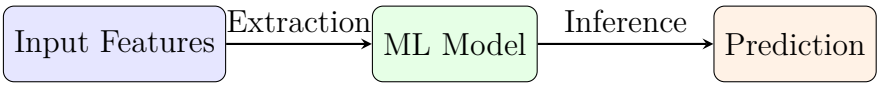


Figure 1.4: Standard supervised learning pipeline.

1.6.2 Learning Paradigms

The application of ML in water pollution detection is generally categorized into several fundamental paradigms based on the availability of labels:

- **Supervised Learning:** This method learns to make predictions from training data that is fully **labeled** [38]. It is mainly used for classification tasks, like identifying types of contamination or Water Quality Classification (WQC), and for regression tasks, like predicting continuous pollutant levels or the Water Quality Index (WQI) [28, 38].

- **Unsupervised Learning:** When data instances lack explicit labels, unsupervised learning algorithms are employed to discover hidden structures within the dataset [38]. In water quality contexts, this is essential for clustering and anomaly detection, allowing systems to identify unusual sensor patterns indicative of sudden contamination events without prior examples [13].

- **Semi-supervised Learning:** Operating between supervised and unsupervised learning, this paradigm utilizes a small amount of labeled data in conjunction with a large pool of unlabeled data. In the water sector, it effectively addresses the scarcity of labels, such as in the development of multi-output soft sensors for wastewater treatment [21].
- **Self-supervised Learning:** Self-supervised frameworks leverage entirely unlabeled data to reconstruct input features and autonomously learn underlying representations. Once pre-trained, these representations can significantly improve downstream predictions, such as chlorophyll-a retrieval from remote sensing imagery in data-scarce inland waters [19].
- **Reinforcement Learning:** This paradigm involves an agent interacting with an environment to maximize cumulative rewards. While possessing strong generalization capabilities, it is seldom applied directly in the field of water environment evaluation compared to the other paradigms [38].

1.6.3 ML Methods Relevant to Water Quality

A diverse array of ML architectures is deployed to handle the specific characteristics of water quality data. For tabular sensor data, classical algorithms such as **Support Vector Machines (SVM)** and decision trees are frequently used [1]. Ensemble methods build on these foundations by combining multiple base models to enhance robustness: **Random Forest (RF)** aggregates the predictions of numerous individual decision trees to reduce generalization error, while **XGBoost** extends sequential boosting with second-order loss optimization and explicit regularization, yielding faster training and improved resistance to overfitting [28]. For image-based tasks, **Convolutional Neural Networks (CNNs)** enable automatic spatial feature extraction from remote-sensing imagery [1]. Table 1.3 provides a structured overview of these methods and their principal characteristics.

Method	**Data Type**	**Task Type**	**Key Strength**
Random Forest	Tabular	Classification & Regression	Robust to overfitting, handles non-linear interactions well.
SVM	Tabular	Classification & Regression	Effective in high-dimensional spaces and robust against outliers.
XGBoost	Tabular	Classification & Regression	High computational efficiency and explicit regularization.
LSTM	Time-series	Forecasting & Anomaly Detection	Captures long-term sequential dependencies.
CNN	Image / Spatial	Classification	Automatic spatial feature extraction (e.g., satellite imagery).

Table 1.3: Comparison of common Machine Learning methods used in Water Quality Detection.

1.6.4 Evaluation Metrics

To rigorously assess the performance of ML models in water quality prediction, specific evaluation metrics are computed based on the nature of the task:

- **Classification Metrics & Confusion Matrix:** Model efficacy is fundamentally measured using a Confusion Matrix, a tabular tool that contrasts predicted classifications against actual labels through four outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [13]. Based on these values, several evaluation metrics are calculated to assess model performance, including Accuracy, Precision, Recall, and the F1-score [28, 13].

  - **Accuracy:** Measures the proportion of correctly classified instances among all predictions.
  $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
  - **Precision:** Evaluates the proportion of predicted positive instances that are actually correct.
  $$Precision = \frac{TP}{TP + FP}$$
  - **Recall:** Measures the ability of the model to correctly identify positive instances.
  $$Recall = \frac{TP}{TP + FN}$$
  - **F1-score:** Represents the harmonic mean of Precision and Recall, providing a balanced evaluation metric.
  $$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

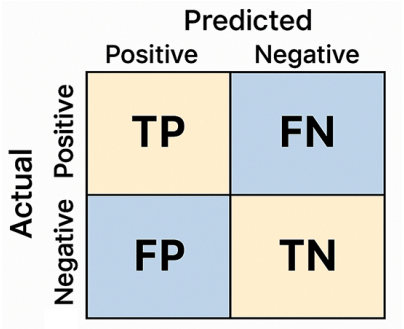


Figure 1.5: Standard binary confusion matrix

- **ROC-AUC:** For binary classification problems—such as specific anomaly detection—the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) provide a comprehensive threshold-independent evaluation of the classifier's performance [13].

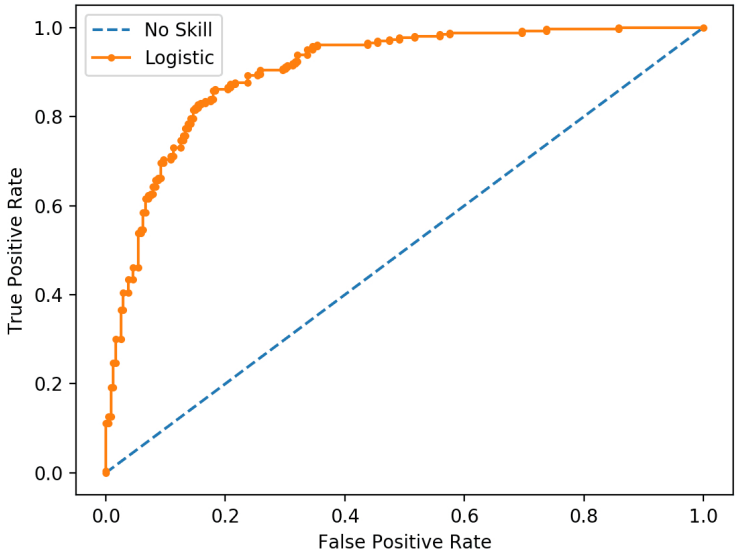


Figure 1.6: Receiver Operating Characteristic (ROC) curve

- **Regression Metrics:** For continuous predictions like the WQI, effectiveness is evaluated using the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE), which quantify the magnitude of prediction errors [28, 38].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Handling Imbalance:** Real-world water quality datasets are inherently imbalanced, as anomalous or polluted events are rare compared to normal conditions. In such scenarios, relying solely on accuracy can be misleading; thus, there is a general consensus in the literature that the F1-score is the preferred performance metric for water quality anomaly detection (WQAD), as it provides a balanced measure of precision and recall [13].

1.7 Conclusion

This chapter presented the theoretical foundations necessary for understanding water pollution detection and intelligent water quality monitoring systems. It first introduced the importance of water resources and the major causes of water pollution, including industrial activities, agricultural runoff, urbanization, and environmental changes. The chapter also discussed the impact of contaminated water on human health, ecosystems, and sustainable development.

Furthermore, the chapter reviewed the main physicochemical and biological parameters used to evaluate water quality, as well as traditional monitoring approaches based

on manual sampling and laboratory analysis. Although these methods provide accurate measurements, they suffer from limitations related to cost, time, and real-time monitoring capability.

In addition, this chapter introduced the fundamental concepts of Artificial Intelligence, Machine Learning, and Deep Learning, highlighting their growing importance in environmental monitoring applications. Different machine learning paradigms, classification techniques, ensemble models, and evaluation metrics were also presented to establish the theoretical basis for the experimental work developed in this thesis. Overall, the chapter demonstrated that intelligent monitoring systems offer promising solutions for improving water quality prediction and pollution detection. The presented concepts and methodologies provide the scientific foundation for the related work analysis and experimental framework discussed in the following chapters.

Chapter 2

Related Work

2.1 Introduction

Machine learning approaches to water-quality monitoring and pollution detection have proliferated in recent years, driven by the increasing availability of physico-chemical sensor data, satellite-derived predictors, and open-access water-quality repositories. The breadth of proposed methods, however, makes it difficult to compare contributions across studies in a principled way: different works rely on different datasets, target different output classes (binary potability, multi-class quality, groundwater potential, anomaly flags), and report their performance through heterogeneous metric sets. This chapter narrows that comparison by structuring related work around the three datasets that frame the experimental contribution of this thesis, denoted *Dataset 1*, *Dataset 2*, and *Dataset 3* throughout. The three

For Dataset 1 and Dataset 2, two recent and methodologically representative works are reviewed in depth (yielding four case studies). For D3, one state-of-the-art work is reviewed, since it is the only directly comparable Dataset 3 benchmark currently registered in the supporting Zotero collection. Each work is described along the same six dimensions—general idea, data, model architecture, results, strengths, and limitations—so that cross-paper comparison is direct. The chapter closes with a synthesis table summarising the principal model type, specific model variant, strengths, limitations, dataset, and evaluation metrics for all five works. The aim is not exhaustive coverage of the water-quality machine-learning literature, but a focused positioning of the present thesis relative to the most directly comparable prior art on the same benchmark datasets.

2.2 Related Works

2.2.1 Dataset 1 – Kaggle Water Quality / Potability Tabular Datasets

Dataset Dataset 1 corresponds to publicly available Kaggle water-quality tabular collections built from physico-chemical and contaminant measurements with a binary potability label. Two recent supervised learning studies on this family of datasets are reviewed below.

Patel et al. (2022) – A Machine Learning-Based Water Potability Prediction Model by Using SMOTE and Explainable AI

Patel et al. [23] proposed a supervised learning pipeline for binary water-potability classification, combining preprocessing (Z-score normalisation, missing-value imputation by feature mean), class-imbalance correction via the **Synthetic Minority Oversampling Technique (SMOTE)**, comparative training of several classical learners, hyperparameter tuning by GridSearchCV and RandomizedSearchCV, and a post-hoc explainability analysis using **Local Interpretable Model-agnostic Explanations (LIME)**.

- **Data:** The authors used the Water Quality dataset from Kaggle, comprising **3,276 samples** described by **nine physico-chemical features**—pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity—with a binary *potability* target. The training set initially contained 1,998 *not potable* and 1,278 *potable* samples, an imbalance that motivated the use of SMOTE to synthesise additional minority-class instances and produce a balanced training set with 1,998 samples per class. The data were then split **70 %/30 %** into training and test partitions [23].

- **Model architecture:** Five supervised classifiers were compared: **Random Forest (RF)**, **XGBoost (XGB)**, Decision Tree (DT), AdaBoost (ADA), and Support Vector Classifier (SVC). Each classifier was first trained with default parameters, then refined through **GridSearchCV** and **RandomizedSearchCV** over algorithm-specific hyperparameter ranges. Repeated stratified k-fold cross-validation was applied to the two best-performing families (RF and XGB) to obtain stable accuracy estimates. The XAI component (LIME) was applied to the final RF model to quantify per-feature contributions to individual predictions [23].

- **Results** After hyperparameter tuning, the Random Forest classifier reached an accuracy of **0.81** with precision, recall, and F1-score all in the 0.80–0.82 range. XGBoost achieved a comparable accuracy of **0.80**, while Decision Tree, AdaBoost, and SVC trailed at 0.73, 0.70, and 0.75 respectively. The LIME analysis identified **sulfate (27 %)**, **pH (18 %)**, and hardness and solids (each at 9 %) as the most influential features for the potability decision, with the remaining contaminants distributed at lower weights [23].

Strengths.

- Explicit treatment of class imbalance via SMOTE, with a before/after sample count reported for full reproducibility.

- Comparative evaluation of fifteen classifiers, with five selected for detailed hyperparameter tuning, reducing the risk of model-selection bias.

- Two complementary hyperparameter-tuning strategies (GridSearchCV and RandomizedSearchCV) applied uniformly across the retained models.

- Post-hoc explainability through LIME, addressing the common transparency limitation of ensemble classifiers.

Limitations.

- The reported accuracy ceiling of $\sim 81\,\%$ is comparatively modest relative to neural-network and hybrid baselines on related Kaggle water-quality variants (see Section 2.2.1 below).
- The feature set is restricted to nine physico-chemical variables; biological and metal-contaminant features available in other Kaggle water-quality variants are not exploited.
- No real-time deployment study or sensor-drift analysis is reported; the evaluation is offline on a static train/test split.

Fatima et al.(2022) – Machine Learning-Based Framework for Water Quality Prediction

Fatima et al. [14] proposed a supervised machine-learning framework for water-quality prediction based on physico-chemical input parameters, with a comparative evaluation of five classical algorithms and a feature-importance analysis using the Random Forest model. The work positions itself as a low-cost, scalable alternative to traditional laboratory-based water-quality assessment.

- **Data:** The study used the Water Quality Prediction dataset from Kaggle, containing **7,999 samples** described by **twenty contaminant features** (aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium) and a binary safety label `is_safe`. The data were preprocessed through mean imputation for missing values, Min–Max normalisation, and Z-score standardisation. An **80 %/20 %** train/test split was used, with correlation-based feature selection retaining variables whose correlation with the binary safety label exceeded a defined threshold [14].
- **Model architecture:** Five supervised algorithms were trained and compared: Linear Regression (LR), Decision Tree (DT), **Random Forest (RF)**, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). All models were implemented in Python using scikit-learn. The Random Forest was retained as the proposed model and trained as an ensemble of decision trees with bootstrap aggregation; predictions were averaged across the trees to produce the final classification. No deep learning, IoT integration, or remote-sensing component is part of the framework [14].
- **Results:** On the held-out test set, Random Forest obtained the best performance with accuracy **92.3 %**, precision 96 %, recall 95 %, and F1-score 96 %. Decision Tree was a close second at **92 %** accuracy, followed by Linear Regression (90.1 %), SVM (90.0 %), and KNN (89.34 %). A baseline comparison reported by the authors positioned the Random Forest result above two prior works on different Kaggle splits, namely a multi-classifier study by Ahmed et al. at 85 % accuracy and a hybrid CEEMDAN-XGBoost / CEEMDAN-RF model by Lu and Ma at 90 % [14].

Strengths.

- Larger feature space (twenty contaminants) than Patel et al. [23], allowing the model to capture metal and biological indicators in addition to standard physico-chemical variables.
- Systematic comparison of parametric, instance-based, kernel-based, and ensemble classifiers on the same data and preprocessing pipeline.
- Explicit reporting of the four classification metrics (accuracy, precision, recall, F1-score), enabling balanced comparison with subsequent works.
- Direct quantitative comparison against prior baseline studies on related Kaggle datasets.

Limitations.

- The work explicitly excludes deep learning and explainable-AI methods, despite their relevance to environmental decision support.
- Class imbalance is not addressed through resampling or cost-sensitive training, although the dataset distribution is not reported as severely skewed.
- No real-time, sensor-stream, or geospatial component is evaluated; the study is restricted to offline tabular prediction.
- Cross-validation results are not reported alongside the train/test evaluation, leaving the variance of the headline accuracy unquantified.

2.2.2 Dataset 2 – Groundwater Quality and Potential Datasets

Dataset Dataset 2 corresponds to groundwater-focused datasets covering both *groundwater quality* (a multi-feature potability label per sample) and *groundwater potential* (a multi-class recharge label per geo-referenced location). The two works reviewed below illustrate the two ends of this spectrum.

Chen et al. (2024) – Machine Learning-Based Water Quality Classification Assessment

Chen et al. [11] addressed groundwater-quality classification through a hybrid model that combines a stacked **Gradient Boosted Decision Tree (GBDT)** feature extractor with a **Multi-Layer Perceptron (MLP)** classifier and a self-constructed **feature-weighted attention** mechanism. The hybrid is named **BS-FAMLP** (Bagging–Stacking Feature-Attention MLP), and its design specifically targets three limitations identified by the authors in single-model approaches: limited individual capacity, the absence of dynamic feature weighting, and information loss from input simplification.

- **Data:** The model was evaluated on a public groundwater-quality dataset released jointly by the European Environment Agency and the World Bank, comprising **188,623 samples** (80,948 labelled *Safe*, 107,675 labelled *Unsafe*) described by **fifteen water-quality parameters**: pH, iron, nitrate, chloride, lead, zinc, turbidity,

fluoride, copper, odour, sulfate, conductivity, chlorine, manganese, and total dissolved solids. The preprocessing pipeline included outlier replacement using a three-IQR criterion on box plots, KNN-based missing-value imputation with $K = 15$ chosen empirically by sweeping forty imputation variants, and Z-score normalisation [11].

- Model architecture. The BS-FAMLP architecture is composed of three stages. The **Bagging layer** contains nine independently trained GBDT models, each fitted on a 12,000-sample bootstrap subset and producing a class-1 probability per input. The **Stacking layer** contains five GBDT models that take the nine probabilities as input features and are trained with five-fold cross-validation. The output of the stacking layer is fed to an MLP with two hidden layers (64 and 128 ReLU neurons), augmented by a feature-weighted attention mechanism that computes per-feature attention weights via a tanh-based scoring function followed by softmax normalisation, with a category-sensitive adjustment that down-weights features whose contributions are similar across classes. Hyperparameters at every stage were optimised through **Bayesian optimisation** [11].
- Results. On the held-out test set, BS-FAMLP achieved an accuracy of **0.9616**, precision of 0.9524, recall of 0.9655, F1-score of 0.9589, and **ROC-AUC of** 0.9834. The ablation study reported by the authors showed that adding the Stacking layer to the Bagging+MLP base improved accuracy from 0.9337 to 0.9423, and that the feature-weighted attention mechanism further raised it to **0.9616**. A multi-head attention variant (BS-MAMLP) reached marginally higher accuracy (0.9632) but at **roughly six times the training cost** [11].

Strengths.

- Very large benchmark dataset (188,623 samples) with a realistic class imbalance, providing a robust testbed for classification models.
- Hybrid GBDT–MLP design that combines tree-based feature extraction with neural non-linear modelling, addressing limitations of either family used alone.
- Self-constructed feature-weighted attention mechanism with an explicit category-sensitive adjustment, dynamically reweighting inputs to mitigate the weak per-feature correlations observed in the data.
- Bayesian hyperparameter optimisation applied uniformly across Bagging, Stacking, and MLP stages.
- Favourable accuracy/efficiency trade-off relative to self-attention and multi-head-attention variants of the same base model.

Limitations.

- The architecture is computationally heavy: training the full BS-FAMLP requires roughly seven times the time of the base Bagging+MLP variant, which may hinder real-time deployment.
- Feature-importance analysis through the attention weights is not externally validated against domain knowledge or a separate XAI method.

- No cross-domain or geographic generalisation experiment is reported; performance is measured on a single shuffled split of the same dataset.
- The work does not address sensor drift, missing-channel robustness, or deployment-time constraints.

Elmotawakkil et al. (2025) – HydroPredictor a Hybrid Machine Learning Model for Addressing Data Scarcity in Groundwater Prediction

E lmotawakkil et al. [13] proposed **HydroPredictor**, a hybrid machine-learning framework for groundwater-potential classification in **data-scarce regions**. The framework integrates a **CatBoost** gradient-boosted tree component that converts raw geo-environmental inputs into structured leaf embeddings, a regularised **Multi-Layer Perceptron (MLP)** that learns non-linear representations from those embeddings, and an ensemble decision-fusion stage (weighted probability fusion or a softmax-based meta-classifier) that combines the two predictors. The explicit design goal is to mitigate overfitting under small-sample training while retaining interpretability.

- Data. The study uses **315 geo-referenced ground-truth samples** collected in the Feija Basin (south-eastern Morocco), evenly stratified across three groundwater-potential classes (*low*, *medium*, *high*; 105 samples each). Each sample is described by **ten environmental predictors**: elevation, soil permeability, annual rainfall, slope, curvature, topographic wetness index, geomorphology, NDVI / land-use–land-cover, stream distance, and lineament density. Continuous variables were normalised by Min–Max scaling, categorical variables one-hot encoded; missing values were imputed by linear interpolation (continuous) or modal imputation (categorical), and Z-score-based outlier filtering was applied. Multicollinearity was controlled through Pearson correlation, mutual-information ranking, and variance-inflation-factor analysis: variables with VIF < 5 were retained without qualification; those with $5 \leq$ VIF < 10 were retained with careful interpretation to avoid discarding environmentally important predictors; and variables with VIF > 10 were excluded due to strong collinearity. The data were split **80 %/20 %** with **five-fold cross-validation** on the training partition [13].
- Model architecture. The HydroPredictor pipeline is structured in three stages. **CatBoost** first produces both a class-probability vector and a tree-leaf embedding $F_{\mathrm{CB}}(x) \in \mathbf{R}^T$ from the raw input, where T is the number of trees. The **MLP** component consists of two ReLU hidden layers, dropout, L_2 regularisation, batch normalisation, and a softmax output, taking the CatBoost embeddings as input. Two fusion strategies are evaluated: a **weighted probability fusion** $P_{\mathrm{final}} = \alpha\, P_{\mathrm{MLP}} + (1-\alpha)\, P_{\mathrm{CB}}$ and a **meta-classifier** (multinomial logistic regression) trained on the concatenated probability vectors. Hyperparameters of all components, including α and the meta-classifier, were tuned through fifty **Optuna** trials with five-fold cross-validation [13].
- Results. On the held-out test set, HydroPredictor attained an accuracy of **89.23 %**, F1-score of 0.8937, recall of 0.8923, **Cohen's κ of** 0.8385, and Matthews correlation coefficient of 0.8394, outperforming standalone CatBoost and Random Forest (both at 84.62 % accuracy), MLP (78.46 %), and SVM (75.38 %). The ROC-AUC values exceeded 0.90 across all three groundwater classes. The Friedman test ($\chi^2 = 12.00$,

$p = 0.0174$) confirmed statistically significant differences among the five models, and the **train/test accuracy gap of** 0.35 % indicates strong generalisation despite the small sample size. A field validation on **90 independent ground-truth points** yielded an overall accuracy of 84.44 % and a Cohen's κ of 0.767 [13].

Strengths.

- Explicit handling of data scarcity ($N = 315$) through regularisation, embedding-based feature transformation, and five-fold cross-validation, with a very small generalisation gap.
- Hybrid tree–neural design that combines categorical-feature handling and ordered boosting (CatBoost) with non-linear representation learning (MLP).
- Bayesian hyperparameter optimisation through Optuna across all stages, replacing exhaustive grid search.
- Two complementary fusion strategies evaluated; statistical significance testing via Friedman and Wilcoxon signed-rank tests.
- Independent field-based validation on 90 ground-truth points, a level of evaluation rarely reported in tabular ML water studies.

Limitations.

- Spatial scope is limited to a single semi-arid Moroccan basin; out-of-region generalisation is not yet demonstrated.
- The hybrid pipeline incurs roughly 45 % more training time than CatBoost alone, a non-trivial cost for resource- constrained settings.
- Validation relies on historical field surveys; real-time integration with satellite (GRACE, MODIS) or IoT telemetry is listed as future work rather than implemented.
- Uncertainty quantification (e.g., Bayesian credible intervals on predictions) is not provided.

2.2.3 Dataset 3 – GECCO 2019 Industrial Challenge Drinking-Water Time Series

Dataset Dataset 3 corresponds to the GECCO 2019 Industrial Challenge drinking-water quality dataset, which provides multivariate time-series sensor readings of water-composition variables with annotated event (anomaly) periods. One recent work that evaluates on this benchmark is reviewed below.

Wang et al. (2025) – FreFilterTST: A Dynamic Channel Graph Sparsification Approach to Multivariate Time Series Anomaly Detection with Frequency-Domain Restoration

Wang, Zhang, and Zhang [33] proposed **FreFilterTST**, an unsupervised multivariate time-series anomaly detection framework that combines **frequency-domain reconstruction** with a **dynamically pruned channel-attention** mechanism. The motivation is

that anomalies of different types (point, contextual, seasonal, trend, shapelet) leave distinguishable signatures in different frequency bands, while standard time-domain detectors often miss high-frequency or subsequence anomalies and either ignore inter-channel correlations (channel-independent methods) or include all of them indiscriminately (channel-dependent methods).

- Data. The framework is benchmarked on six public datasets, one of which is directly relevant to this thesis: the **GECCO 2019 Industrial Challenge** drinking-water composition dataset, in which the task is to detect events in a multivariate stream of drinking-water sensor readings. The remaining five datasets (Calit2, MSL, NYC, SEU, PSM) are non-water benchmarks used to demonstrate cross-domain robustness; only the GECCO results are retained for comparison here. Inputs are normalised through **Reversible Instance Normalisation (RevIN)** prior to the Fourier transform [33].

- Model architecture. The input multivariate series $X \in \mathbf{R}^{N \times T}$ is first transformed by a **Fast Fourier Transform**; the real and imaginary spectra are then segmented into n non-overlapping *frequency patches* of length p, projected into a higher dimensional embedding through a linear-plus-GELU layer, and passed to the core *FreFilter* module. FreFilter combines a **Frequency Channel Filter (FCF)** with a **Channel Transformer (CT)**. The FCF builds a dense channel-similarity adjacency matrix, applies a top-k structural sparsification, and dynamically combines two canonical expert masks—an Amplitude Expert (fully connected) and a Frequency Expert (identity)—through a **noisy top-gating Mixture-of-Experts**, producing a sparse attention mask that guides the Channel Transformer's cross-channel attention. The output is reconstructed back to the time domain through an **inverse FFT**, and the anomaly score combines a time-domain MSE term and a frequency-domain MAE term on the amplitude spectrum [33].

- Results. On the GECCO 2019 benchmark, FreFilterTST achieved an Affiliation-based F1-score of **0.912** and a **ROC-AUC of 0.966**, outperforming twelve baselines that span classical methods (Isolation Forest, PCA, HBOS), linear models (DLinear, NLinear), the autoencoder, and state-of-the-art deep learners (PatchTST, TimesNet, ModernTCN, DCdetector, TFAD, DualTF). The closest deep-learning competitors on GECCO were ModernTCN (AUC 0.952) and PatchTST (AUC 0.957). The ablation study reported by the authors confirmed that both expert masks contribute to the final score, and the model is reported to be highly compact: **2.8 M parameters** and **0.12 GFLOPs** per forward pass on GECCO, against ≈ 211 M parameters and 1.94 GFLOPs for the channel-clustering baseline [33].

Strengths.

- Unsupervised formulation, removing the requirement for labelled anomaly events at training time.

- Frequency-domain patching enables fine-grained spectral analysis and explicit detection of high-frequency events that purely time-domain detectors miss.

- Adaptive Mixture-of-Experts gating dynamically prunes spurious channel correlations, addressing the capacity/robustness trade-off between channel-independent and channel-dependent approaches.

- Very low parameter count and GFLOP cost relative to competing deep models, supporting deployment on edge or embedded sensor platforms.
- Dual time-and-frequency anomaly score with a weighting hyperparameter λ allows tuning to the dominant anomaly type expected in deployment.

Limitations.

- The framework is reconstruction-based: anomalies characterised by a sustained *absence* of signal (e.g., the Calit2 case study reported by the authors) yield low reconstruction error and may be missed.
- The FFT-based representation assumes stationarity within each analysis window, which can blur transient or non-stationary spectral events.
- The dense channel-similarity step has $O(N^2)$ memory cost, limiting scalability to ultra-high-dimensional sensor systems ($N > 1{,}000$); the authors discuss approximate-nearest-neighbour extensions as future work.
- Only the GECCO benchmark among the six evaluated datasets is a water-domain dataset; cross-domain transferability to other water-quality streams is not directly demonstrated.

2.3 Comparative Synthesis

Research & Year	**Principal Model Type**	**Specific Model**	**Strengths**	**Limitations**	**Dataset**	**Metrics**
Patel et al. (2022) [23]	Ensemble (tree-based)	Random Forest + SMOTE + LIME	– SMOTE rebalancing – Compared 15 classifiers – LIME XAI	– Modest acc. (∼81%) – 9-feature input – No streaming evaluation	Kaggle Water Potability (D1; 3,276 samples, 9 features)	Acc. 0.81 (RF) 0.80 (XGB) Prec. 0.83 Recall 0.80 F1 0.82
Fatima et al. [14]	Ensemble (tree-based)	Random Forest (vs. DT, SVM, KNN, LR)	– 20-feature dataset – 5-classifier benchmark – Acc/Prec/Rec/F1 reported	– No deep learning, no XAI – No imbalance handling – No CV variance reported	Kaggle Water Quality Pred. (D1; 7,999 samples, 20 features)	Acc. 92.3% Prec. 96% Recall 95% F1 96%
Chen et al. (2024) [12]	Hybrid (boosting + neural + attention)	BS-FAMLP (Bag.–Stack. GBDT → MLP + feat.-weighted attention)	– 188K-sample benchmark – Hybrid GBDT+MLP – Dynamic feature attention – Bayesian HPO	– ∼7× slower than base – Attention not externally validated – Single-split evaluation	EEA / World Bank groundwater (D2; 188,623 samples, 15 features)	Acc. 0.9616 Prec. 0.9524 Recall 0.9655 F1 0.9589 AUC 0.9834
Elmotawakkil et al. (2025) [13]	Hybrid (boosting + neural + ensemble fusion)	HydroPredictor (CatBoost embed. → MLP + meta-classifier)	– Robust at $N = 315$ – Optuna HPO – Friedman / Wilcoxon tests – Field validation, 90 sites	– Single-basin scope – ∼45% extra train time – No real-time integration – No uncertainty estimates	Feija Basin, Morocco (D2; 315 geo-ref. samples, 10 predictors, 3 classes)	Acc. 0.8923 F1 0.8937 Recall 0.8923 κ 0.8385 MCC 0.8394 AUC > 0.90
Wang et al. (2025) [33]	Deep learning (freq.-domain Transformer + MoE)	FreFilterTST (FFT patches → MoE-gated Channel Transformer → iFFT)	– Unsupervised – Spectral patches – Adaptive MoE pruning – 2.8M params, 0.12 GFLOPs	– Misses signal-absence anomalies – Stationarity assumption – $O(N^2)$ adjacency memory	GECCO 2019 Industrial Challenge drinking water (D3)	Aff-F1 0.912 AUC-ROC 0.966 (GECCO only)

Table 2.1: Comparative summary of the five related works grouped by dataset (D1, D2, D3).

2.4 Comparative Synthesis

The reviewed studies demonstrate the rapid advancement of machine learning methods for water quality prediction, groundwater assessment, and anomaly detection. Despite differences in datasets and methodologies, several common trends and research gaps emerge across Dataset 1, Dataset 2, and Dataset 3.

2.4.1 Dataset 1 - Tabular Water Potability Classification

- Ensemble tree models such as Random Forest and XGBoost remain the dominant approaches for tabular water quality prediction.
- **Patel et al. (2022)** [23] used SMOTE with Random Forest and achieved approximately 81% accuracy, while **Fatima et al.** achieved 92.3% using a richer 20-feature dataset.
- Results indicate that feature richness and feature engineering strongly influence performance, sometimes more than model selection itself.
- Both studies rely on traditional preprocessing pipelines including imputation and normalization.
- Explainability remains limited, with only Patel et al. applying LIME-based interpretation.
- Neither study investigates threshold calibration, uncertainty estimation, or temporal generalization.
- The effect of synthetic oversampling on engineered environmental features remains unexplored.

Research Gap 1. *No previous Dataset 1 study systematically compares multiple resampling strategies with optimized ensemble methods under engineered feature spaces.*

2.4.2 Dataset 2 - Groundwater Quality and Potential Prediction

- Recent Dataset 2 studies converge toward hybrid boosting–neural architectures.
- **Chen et al.** [12] combined GBDT, attention mechanisms, and MLPs to achieve very high accuracy on a large dataset.
- **Elmotawakkil et al.** [14] demonstrated that hybrid CatBoost–MLP models can generalize effectively even with only 315 samples.
- Bayesian optimization is increasingly adopted for hyperparameter tuning.
- Hybrid models successfully capture complex non-linear environmental relationships.
- However, computational complexity and deployment cost increase significantly.
- Spatial transferability, uncertainty estimation, and real-time deployment are still insufficiently addressed.

- The interaction between severe imbalance and SMOTE-based augmentation is not experimentally studied.

Research Gap 2. *No Dataset 2 work evaluates whether synthetic oversampling may degrade performance under extreme imbalance conditions.*

2.4.3 Dataset 3 - Time-Series Anomaly Detection

- **Wang et al.** [35] introduced *FreFilterTST*, a frequency-domain Transformer framework for unsupervised anomaly detection.
- The model achieved state-of-the-art performance on the GECCO 2019 drinking-water benchmark.
- Frequency-domain analysis improves detection of spectral anomalies and temporal irregularities.
- The framework combines FFT processing, Mixture-of-Experts gating, and Transformer attention.
- Despite strong performance, the method assumes local stationarity and has scalability limitations due to high memory complexity.
- Reconstruction-based anomaly detection may fail for signal-absence anomalies.
- Comparisons against engineered tree-based ensemble approaches remain absent.

Research Gap 3. *No Dataset 3 study demonstrates whether engineered temporal features with optimized tree ensembles can outperform deep learning architectures.*

2.4.4 Cross-Dataset Methodological Trends

Hybrid Architectures

- Recent studies increasingly combine boosting, neural networks, attention mechanisms, and ensemble fusion.
- Hybrid systems aim to capture both structured and non-linear environmental relationships.

Inconsistent Imbalance Handling

- Some works apply SMOTE, others use class weighting, while several ignore imbalance entirely.
- There is no unified guideline regarding when resampling is beneficial or harmful.

Growing Use of Bayesian Optimization

- Bayesian optimization is gradually replacing GridSearch and RandomizedSearch due to higher efficiency.
- However, tuning strategies remain inconsistent across studies.

Limited Explainability

- Most studies prioritize predictive performance over environmental interpretability.
- Explainable AI remains underutilized in water quality prediction systems.

2.4.5 Persistent Research Gaps

The literature reveals several unresolved issues:

- Lack of systematic comparison between resampling and native imbalance handling.
- Limited reporting of reproducible feature engineering strategies.
- Absence of studies investigating SMOTE artifacts in engineered feature spaces.
- Neglect of threshold calibration for imbalanced datasets.
- Insufficient reporting of computational cost and deployment feasibility.

2.4.6 Positioning of the Present Thesis

The present thesis addresses these gaps through:

1. A dual-phase framework comparing raw imbalanced training with multiple resampling methods.
2. Extensive domain-informed feature engineering.
3. Consistent Bayesian optimization using *Optuna*.
4. Threshold calibration for all classification models.
5. Evaluation of computational efficiency alongside predictive performance.
6. Demonstration that optimized ensemble methods with engineered features can rival or exceed complex deep-learning systems.

2.5 Conclusion

The reviewed studies establish strong baselines for water-quality machine learning using ensemble learning, hybrid architectures, and deep temporal models. However, important methodological limitations remain unresolved, particularly regarding imbalance handling, feature engineering transparency, explainability, and deployment realism. The present thesis contributes by providing a unified and experimentally rigorous framework across three datasets, while demonstrating that carefully engineered features and optimized

ensemble models can achieve state-of-the-art environmental prediction performance without excessive architectural complexity.

Chapter 3

Methodology and Data Exploration

3.1 General Methodology Overview

This section outlines the foundational framework applied to all three datasets to ensure consistency, rigor, and reproducibility in the water pollution detection experiments.

3.1.1 Experimental Pipeline: From Preprocessing to Evaluation

The experimental workflow follows a modular pipeline designed to transform raw environmental sensor data into actionable insights through four primary stages:

- **Data Preparation and Initial Cleaning:** This stage involves processing the three datasets by resolving missing values and logical inconsistencies. while systematic exploration using correlation heatmaps and boxplots identifies pollution patterns and feature dependencies.
- **Feature Engineering and Pre-Processing:** Feature engineering adds predictive criteria, account for varying units, numerical features are standardized. This prevents features with larger magnitudes from disproportionately influencing model performance, ensuring uniform scaling across all physical and chemical parameters.
- **The Dual-Phase Training Engine:** Models are trained in two distinct stages. Phase 1 utilizes the original imbalanced data to establish a baseline, while Phase 2 integrates targeted imbalance mitigation techniques to address class skewness, followed by robust performance validation.
- **Ensemble Integration & Evaluation:** Predictive power is maximized by combining multiple models via **Stacking** (using a meta-learner) and **Soft Voting** (averaging class probabilities). The final model is calibrated through **Decision Threshold Optimization** to maximize the F1-score, with reliability confirmed via precision-recall curves and confusion matrices.

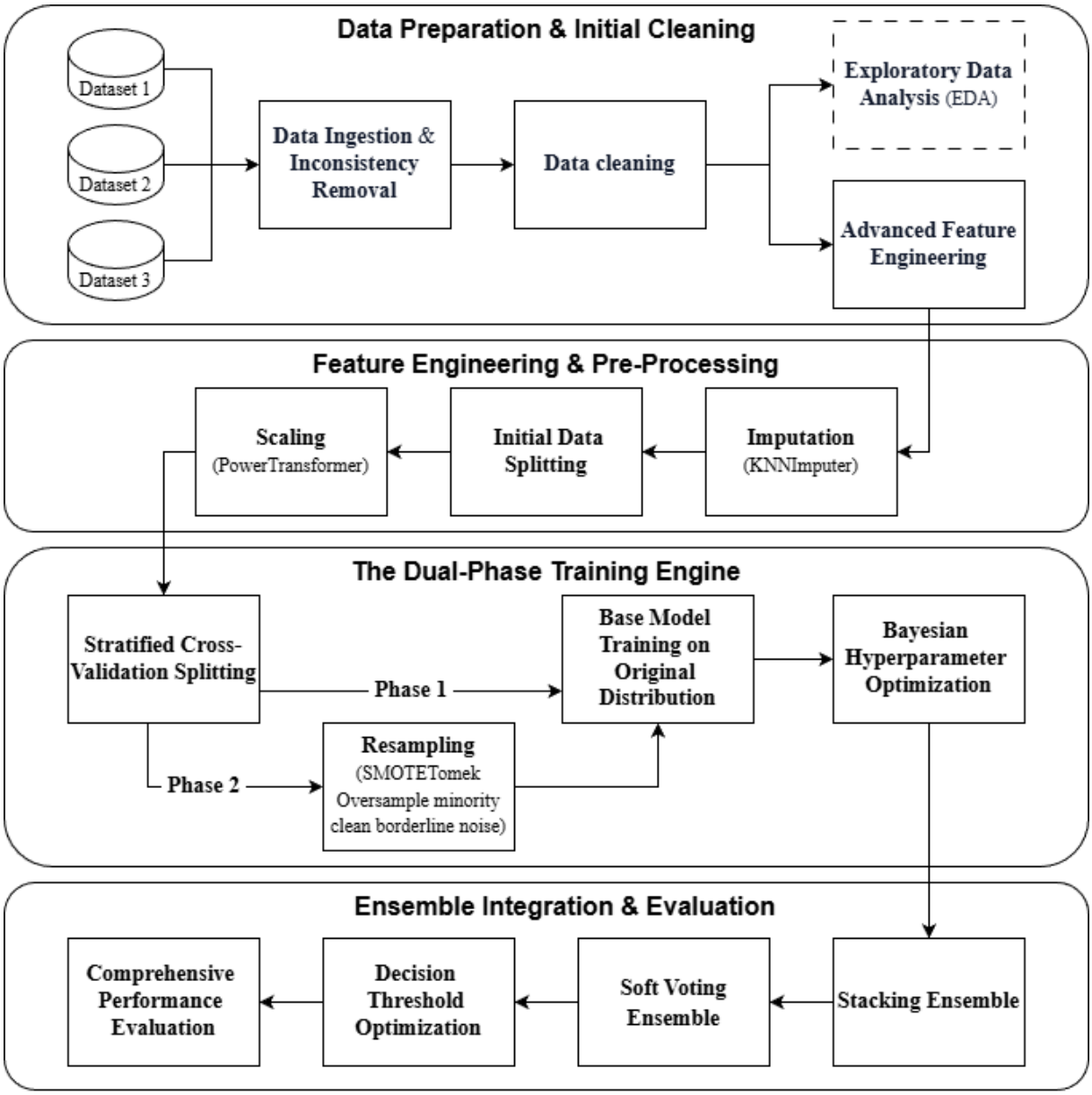


Figure 3.1: Workflow Pipeline Diagram

3.1.2 Feature Exploration Methods

Distribution Per Target

This method visualizes how independent features behave across different target classes. By comparing these distributions, features with high predictive power can be identified; significant separation between classes suggests a feature is a strong candidate for model training.

Histograms & KDE Plots

Histograms count frequencies within data bins, while Kernel Density Estimate (KDE) plots provide a smoothed probability density of continuous variables. These reveal the

data's "shape," including skewness (lean), modality (peaks indicating sub-groups), and outliers.

The KDE function can be expressed as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where n is the number of samples, h is the bandwidth parameter, K is the kernel function, and x_i represents individual observations.

Correlation Heatmap

This graphical representation uses a color scale to show relationship strengths between all numerical variables. It is essential for feature selection and identifying multicollinearity; features with a correlation near 1.0 may be removed to simplify the model and prevent overfitting.

The Pearson correlation coefficient is commonly used and is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i and y_i are sample values, while $\bar{x}$ and $\bar{y}$ denote their respective means.

Boxplots by Event

Boxplots display a five-number summary (minimum, Q1, median, Q3, and maximum). Categorizing them "by event" allows for a side-by-side comparison of central tendencies and spreads, making it the standard tool for identifying outliers and comparing data distributions across categories.

The Interquartile Range (IQR), used for outlier detection in boxplots, is calculated as:

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 represent the first and third quartiles.

Pair Plots

A Pair Plot generates a grid of scatter plots for every variable pair, typically featuring histograms or KDEs on the diagonal. This provides a comprehensive overview of the dataset, helping to uncover non-linear relationships and clusters that individual correlation coefficients might overlook.

3.1.3 Methodological Framework and Pipeline Architecture

Imputation (KNNImputer)

Standard imputation methods often rely on simple metrics like mean or median. The **KNNImputer** leverages the k-Nearest Neighbors algorithm to identify the k most similar samples based on available features. By averaging values from these neighbors, it preserves the local correlation structure of the data, leading to more realistic data. The similarity between samples is commonly measured using the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where x_i and y_i represent feature values of two samples.

Initial Data Splitting

The first step in the validation strategy is the **Train-Test Split**. A fixed portion of the data (20% in this study) is completely isolated from the training process. This "hold-out" set serves as the final arbiter of performance, ensuring the model is evaluated on entirely unseen data to prevent data leakage and over-optimistic results.

Scaling (PowerTransformer / Yeo-Johnson)

While standard scaling adjusts mean and variance, many algorithms perform optimally only when features follow a Gaussian distribution. The **Yeo-Johnson Power Transformation** is applied to minimize skewness and stabilize variance across features. Unlike other methods, it supports both positive and negative values, making feature distributions more bell-shaped and helping learners converge more accurately. The Yeo-Johnson transformation is defined as:

$$Y(\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0,\ y \geq 0 \\ \ln(y+1), & \lambda = 0,\ y \geq 0 \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2,\ y < 0 \\ -\ln(-y+1), & \lambda = 2,\ y < 0 \end{cases}$$

Stratified Cross-Validation

In imbalanced datasets where the target class is rare, a random split can lead to training folds that lack minority examples. **Stratification** ensures that each cross-validation fold maintains the exact class proportions of the original dataset. Utilizing **Stratified 5-Fold CV** forces the model to learn from the minority class in every iteration, providing a stable and reliable estimate of predictive performance. Stratification preserves the original class distribution in each fold:

$$\frac{n_c^{(k)}}{N^{(k)}} \approx \frac{n_c}{N}$$

where $n_c^{(k)}$ is the number of samples of class c in fold k, and $N^{(k)}$ is the total number of samples in that fold.

Resampling Techniques

To address class imbalance and improve the representation of minority classes, several resampling techniques are explored:

- **SMOTE:** Generates synthetic samples for the minority class by interpolating between neighboring minority instances, helping to balance the dataset without simple duplication. Synthetic samples in SMOTE are generated using interpolation:

$$x_{new} = x_i + \lambda(x_{nn} - x_i)$$

where x_i is a minority sample, x_{nn} is one of its nearest neighbors, and $\lambda \in [0, 1]$ is a random value.

- **BorderlineSMOTE:** A variation of SMOTE that focuses on generating synthetic samples near the decision boundary, where minority instances are more likely to be misclassified. Focuses synthetic generation on minority samples located near the decision boundary. Samples are selected from the "danger" region:

$$D(x_i) = \frac{\text{Majority Neighbors}}{k}$$

where $D(x_i)$ measures the proportion of majority-class neighbors among the k-nearest neighbors.

- **SMOTE-ENN:** Combines SMOTE oversampling with Edited Nearest Neighbor (ENN) cleaning. A sample is removed if its class label differs from the majority of its neighbors:

$$y_i \neq \text{mode}(y_{NN})$$

where y_{NN} represents the labels of neighboring samples.

- **SMOTETomek:** Integrates SMOTE with Tomek Links undersampling. After generating synthetic minority samples, Tomek Links are removed to eliminate overlapping and ambiguous samples near class boundaries, resulting in a cleaner and more balanced dataset. Integrates SMOTE with Tomek Links undersampling. Two samples form a Tomek Link if:

$$d(x_i, x_j) < d(x_i, x_k)$$

for every other sample x_k, where x_i and x_j belong to different classes. Such pairs are removed to reduce class overlap.

Bayesian Hyperparameter Optimization

To avoid the inefficiency of Grid or Random search, the pipeline utilizes **Bayesian Optimization** via the Tree-structured Parzen Estimator (TPE). This methodology builds a probability model of the objective function, using the results of previous trials to intelligently select the next set of hyperparameters (such as learning rate or tree depth) to test. This process identifies optimal configurations with high mathematical precision and significantly less computational overhead. Bayesian Optimization seeks the optimal hyperparameters x^* by maximizing an objective function:

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

where $f(x)$ represents the model performance for a given set of hyperparameters.

Stacking Ensemble

This approach utilizes a two-layer architecture to minimize individual model biases. Level-0 consists of diverse base learners, while their outputs are used as input features for a Level-1 meta-learner (typically Logistic Regression), which learns how to combine them optimally.

Let M denote the number of base models. Each base model produces a prediction:

$$h_m(x), \quad m = 1, 2, \ldots, M$$

These predictions are concatenated into a new feature vector:

$$z = [h_1(x), h_2(x), \ldots, h_M(x)]$$

The meta-learner then produces the final prediction:

$$\hat{y} = g(z) = g(h_1(x), h_2(x), \ldots, h_M(x))$$

where $g(\cdot)$ is the learning function of the meta-model.
In probabilistic classification, stacking can also be written as:

$$\hat{y} = g\left(P_1(y|x), P_2(y|x), \ldots, P_M(y|x)\right)$$

where $P_m(y|x)$ represents the predicted probability from each base model.

Soft Voting Ensemble

Unlike hard voting, which relies on a majority count of labels, **Soft Voting** averages the predicted probabilities across all models in the ensemble. This gives more weight to high-confidence predictions; for instance, a single model with high certainty can outweigh several models with marginal certainty, leading to more robust decision-making in high-precision scenarios. The final prediction probability in soft voting is computed as:

$$P(y = c) = \frac{1}{M} \sum_{m=1}^{M} P_m(y = c)$$

where M is the number of models and $P_m(y = c)$ is the predicted probability of class c from model m.

Decision Threshold Optimization

Standard classifiers use a default probability threshold of 0.5 to assign labels. However, in critical domains like water safety, the cost of a missed detection (False Negative) is significantly higher than a False Positive. This phase involves sweeping a range of values (e.g., 0.2 to 0.8) to find the specific threshold that maximizes the **F1-score**, ensuring the model's sensitivity is perfectly calibrated to the specific requirements of the task. Predicted labels are assigned based on a probability threshold t:

$$\hat{y} = \begin{cases} 1, & P(y = 1) \geq t \\ 0, & P(y = 1) < t \end{cases}$$

3.1.4 Feature Engineering Methods

Let $\mathbf{x}$ denote a generic feature vector and $\mathbf{x}_t$ its temporal extension. Across all datasets, we design a unified feature engineering framework combining physicochemical transformations, domain-specific composite indices, and temporal signal modeling. The objective is to move from raw measurements to a richer representation $\mathbb{R}^d \to \mathbb{R}^{d'}$ that captures nonlinear interactions, aggregated pollution burden, and dynamic behavior [16, 18].

Physicochemical Feature Transformations

To model nonlinear relationships between contaminants, we define the following transformations:

Pairwise Ratio Features

$$r_{a,b} = \frac{x_a}{x_b + \varepsilon}, \tag{3.1}$$

capturing relative contamination levels between interacting substances [16].

Interaction Features

$$x_{a,b}^{(int)} = x_a \cdot x_b, \tag{3.2}$$

used to model synergistic pollutant effects [16].

Polynomial Transformations

$$x_i^{(2)} = x_i^2, \tag{3.3}$$

to capture nonlinear toxicity amplification [16].

Logarithmic Stabilization

$$x_i^{(\log)} = \log(1 + |x_i|), \tag{3.4}$$

applied to skewed distributions with heavy tails [6, 31].

Domain-Informed Composite Indices

To incorporate prior knowledge about pollutant groups, we define aggregated indices:

Heavy Metal Index

$$H_{\text{metal}} = \sum_{i \in \mathcal{H}} x_i, \quad H_{\max} = \max_{i \in \mathcal{H}} x_i, \tag{3.5}$$

motivated by environmental pollution aggregation studies [32, 20].

Organic Load Index

$$O_{\text{load}} = \sum_{i \in \mathcal{O}} x_i, \tag{3.6}$$

Physical Condition Index

$$P_{\text{phys}} = \sum_{i \in \mathcal{P}} x_i, \tag{3.7}$$

where $\mathcal{H}, \mathcal{O}, \mathcal{P}$ represent heavy metals, organic pollutants, and physical parameters respectively [32].

Statistical Aggregates

To summarize global sample behavior:

$$\mu = \frac{1}{d}\sum_{i=1}^{d} x_i, \tag{3.8}$$

$$\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2}, \tag{3.9}$$

$$x_{\max} = \max(x_1, \ldots, x_d), \tag{3.10}$$

$$x_{\min} = \min(x_1, \ldots, x_d), \tag{3.11}$$

$$\text{range} = x_{\max} - x_{\min}. \tag{3.12}$$

These descriptors follow standard statistical learning formulations [18].

Temporal Feature Engineering

For sensor streams $\mathbf{x}_t$, we model dynamics using rolling statistics:

Rolling Statistics

$$\mu_t^{(i,w)} = \frac{1}{w}\sum_{k=0}^{w-1} x_{t-k}^{(i)}, \tag{3.13}$$

$$\sigma_t^{(i,w)} = \sqrt{\frac{1}{w}\sum_{k=0}^{w-1}(x_{t-k}^{(i)} - \mu_t^{(i,w)})^2}. \tag{3.14}$$

standard in time-series feature extraction [27, 30].

Temporal Differences

$$\Delta x_t^{(i)} = x_t^{(i)} - x_{t-1}^{(i)}, \tag{3.15}$$

$$\Delta^2 x_t^{(i)} = x_t^{(i)} - x_{t-2}^{(i)}. \tag{3.16}$$

used to capture trends and change rates in time series data [27].

Temporal Interaction Terms

$$x_t^{(a\times b)} = x_t^{(a)} \cdot x_t^{(b)}. \tag{3.17}$$

Anomaly Detection Features

Standardized anomaly scores are defined as:

$$z_t^{(i)} = \frac{x_t^{(i)} - \mu^{(i)}}{\sigma^{(i)}}, \tag{3.18}$$

with binary anomaly indicators:

$$a_t^{(i)} = \begin{cases} 1 & \text{if } |z_t^{(i)}| > 3, \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

The global anomaly signal is:

$$A_t = \sum_{i=1}^{d} a_t^{(i)}. \tag{3.20}$$

Anomaly detection via z-score thresholding is a standard statistical monitoring technique [22, 2].

3.2 Dataset 1: Urban Water Quality (Kaggle)

3.2.1 Dataset Description

The first dataset is sourced from the Kaggle repository and contains 7,996 records of water samples. Each sample is assessed based on the concentration of 20 distinct chemical and biological attributes. This dataset has been used in prior machine-learning studies on water quality classification [3, 29, 15].[1]

The features encompass a wide range of **chemical, biological, and radiological attributes**, including heavy metals (Arsenic, Lead, Mercury), chemical compounds (Nitrates, Silver, Chloramine), and biological indicators. These represent the diverse nature of potential pollutants in a city's water infrastructure.

Regarding **target labeling**, the task is a binary classification in which the target variable is defined as "Safe" (1) or "Not Safe" (0). Safety is determined by whether the concentrations of the 20 substances fall within the permissible limits set for human consumption.

[1]The dataset is publicly available on Kaggle: `https://www.kaggle.com/datasets/mssmartypants/water-quality`

Parameter	Type	Hazard Threshold
Aluminium	Chemical	> 2.8 mg/L
Ammonia	Chemical	> 32.5 mg/L
Arsenic	Chemical	> 0.01 mg/L
Barium	Chemical	> 2 mg/L
Cadmium	Chemical	> 0.005 mg/L
Chloramine	Chemical	> 4 mg/L
Chromium	Chemical	> 0.1 mg/L
Copper	Chemical	> 1.3 mg/L
Fluoride	Chemical	> 1.5 mg/L
Bacteria	Biological	> 0 CFU/mL
Viruses	Biological	> 0 PFU/mL
Lead	Chemical	> 0.015 mg/L
Nitrates	Chemical	> 10 mg/L
Nitrites	Chemical	> 1 mg/L
Mercury	Chemical	> 0.002 mg/L
Perchlorate	Chemical	> 56 mg/L
Radium	Radiological	> 5 mg/L
Selenium	Chemical	> 0.5 mg/L
Silver	Chemical	> 0.1 mg/L
Uranium	Radiological	> 0.3 mg/L

Table 3.1: Urban Water Quality Parameters and Hazard Thresholds (adapted from [36])

3.2.2 Exploratory Data Analysis (EDA)

Water Quality Distribution per Target

The dataset exhibits a significant class imbalance. Approximately 88.6% of the samples are classified as "Not Safe," while only 11.4% are "Safe." This skew reflects the stringent nature of safety standards—where the presence of a single contaminant above the threshold renders the entire sample unsafe.

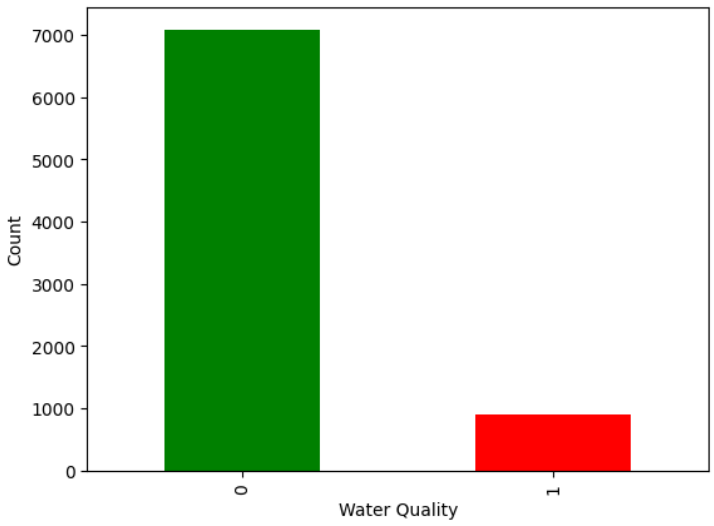


Figure 3.2: Dataset 1 Water Quality Distribution per Target

Histograms & KDE Plots

Density plots reveal that most contaminants follow a right-skewed distribution. A large majority of samples contain very low concentrations of toxic substances, with a small number of "outlier" samples showing high spikes that trigger the "Not Safe" label.



Figure 3.3: Dataset 1 Histograms & KDE Plots

Correlation Heatmap

The heatmap indicates low overall correlation between the different chemical features. This suggests that the presence of one contaminant (e.g., Lead) does not necessarily imply the presence of another (e.g., Nitrates), requiring the model to evaluate all 20 features independently.

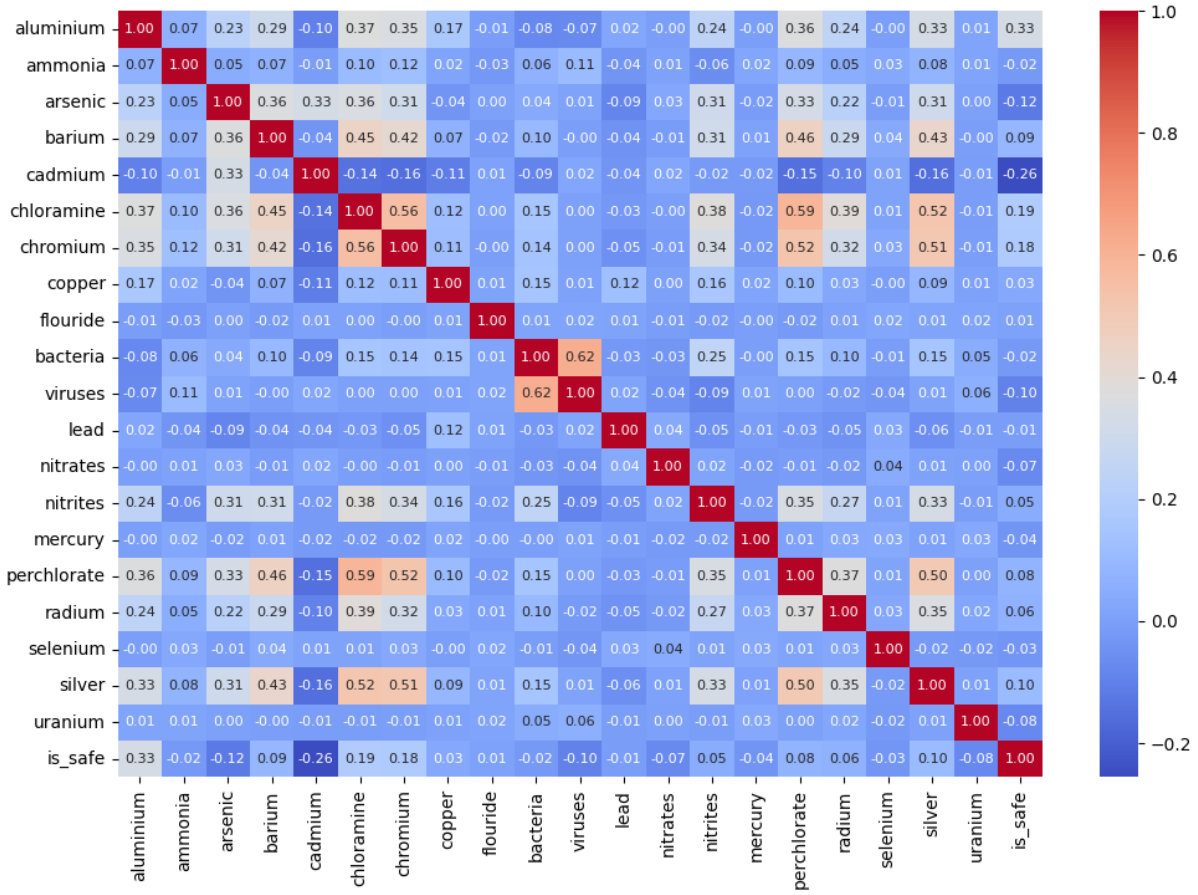


Figure 3.4: Dataset 1 Correlation Heatmap

Boxplots by Event

Boxplots grouped by the target class highlight that "Not Safe" samples often contain extreme outliers in specific categories like Cadmium or Perchlorate. These visualizations help confirm which features provide the strongest discriminative power between the two classes.

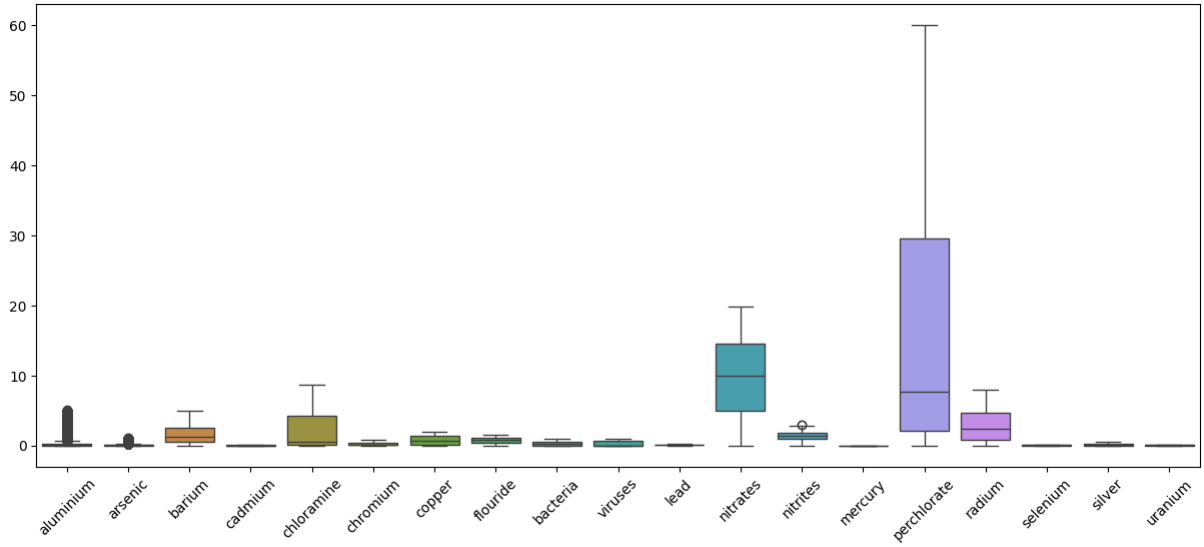


Figure 3.5: Dataset 1 Boxplots by Event

Pair Plots

Pair plots demonstrate that there is no simple linear boundary that separates safe water from unsafe water. Instead, the safety zone is a tight cluster in high-dimensional space, justifying the use of non-linear models like Random Forests or SVMs.

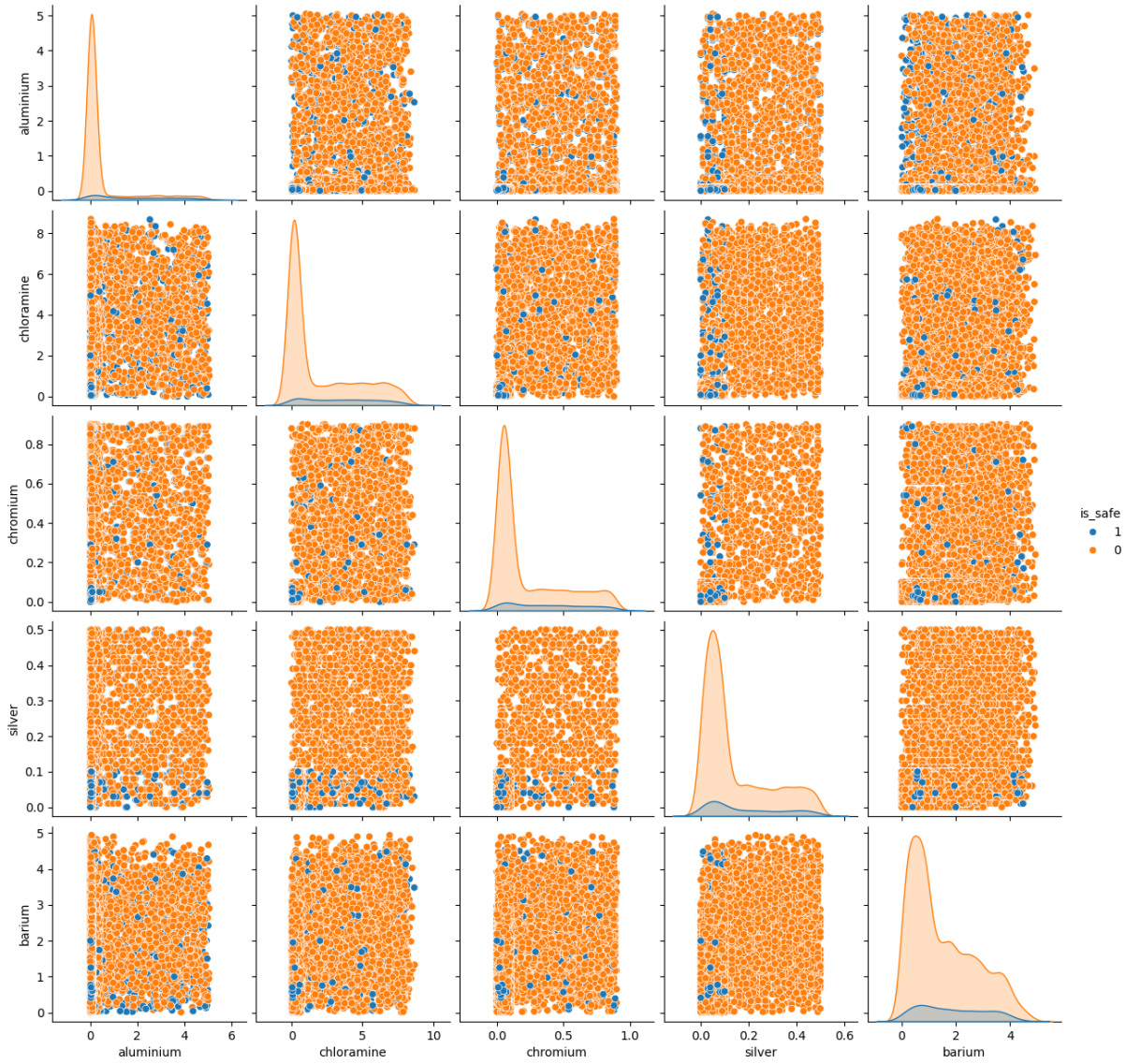


Figure 3.6: Dataset 1 Pair Plots

3.3 Dataset 2: Physicochemical Pollution

3.3.1 Dataset Description

The second dataset, authored by Al-Madhoun and Abu-Naser [3], provides a more granular look at water quality through a multiclass lens. It consists of samples characterized by physical and chemical parameters collected from various water sources.

Parameter	Unit	Min	Max	Description
pH	—	6.5	8.5	Measures acidity or alkalinity; extreme levels can cause pipe corrosion or heavy metal leaching.
Turbidity	NTU	0	1	Measures water clarity; high levels indicate suspended solids that can shield pathogens from disinfection.
Temperature	°C	0	25	Influences chemical reaction rates and biological growth; cooler water is generally preferred for taste.
Dissolved Oxygen (DO)	mg/L	6	—	Indicates the amount of free oxygen available for aquatic life; low levels often signal high organic pollution.
BOD	mg/L	0	5	Measures the amount of oxygen required by bacteria to break down organic matter in the water.
Lead (Pb)	mg/L	0	0.01	A toxic heavy metal that accumulates in the body; primarily enters water through aging plumbing systems.
Mercury	mg/L	0	0.001	A highly toxic heavy metal that can cause neurological damage; often enters water through industrial waste or natural deposits.
Arsenic	mg/L	0	0.01	A natural semi-metallic element that is highly toxic and carcinogenic; commonly found in groundwater in certain geological regions.

Table 3.2: Physicochemical Pollution Dataset Parameters

- **Multiclass Categorization of Pollution Levels:** Unlike the first dataset, this study categorizes water into multiple tiers (e.g., Clean, Slightly Polluted, Moderately Polluted, and Heavily Polluted). This allows the model to predict the severity of contamination rather than a simple binary status.
- **Regulatory Compliance Standards (WHO/EPA):** The labeling of this dataset is grounded in international water quality standards, such as those provided by the World Health Organization (WHO) [36] and the Environmental Protection Agency (EPA). These standards define the specific ranges for pH, turbidity, dissolved oxygen, and other minerals that determine the classification of each sample.

3.3.2 Exploratory Data Analysis (EDA)

The EDA for Dataset 2 focuses on how different pollution levels manifest across physical and chemical dimensions.

Water Quality Distribution per Target

The distribution analysis reveals the frequency of each pollution tier, highlighting any existing class imbalances within the physicochemical samples. In environmental monitoring, "Clean" water often dominates the dataset, while "Heavily Polluted" instances are statistically rarer but critically important to identify. Documenting this baseline distribution justifies the subsequent use of balancing techniques to ensure the model does not become biased toward the majority class.

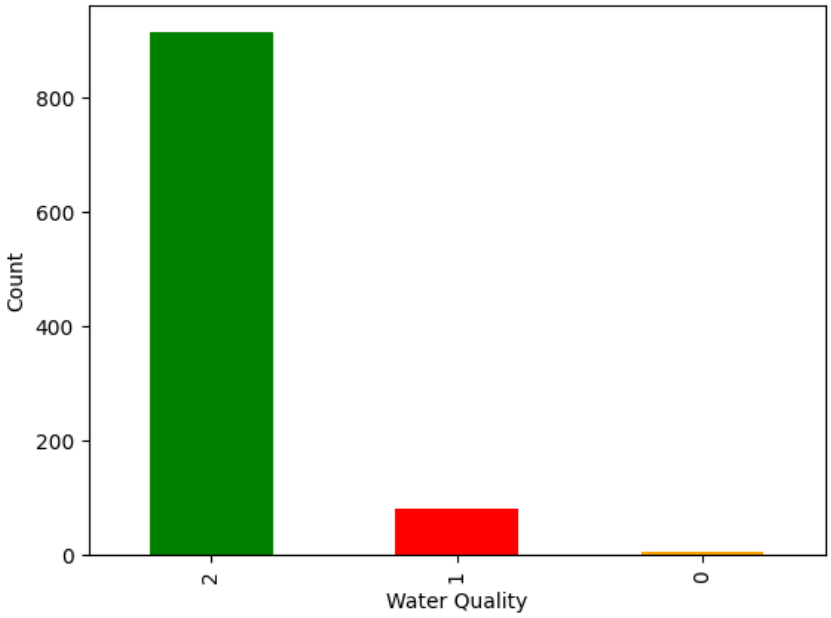


Figure 3.7: Dataset 2 Water Quality Distribution per Target

Histograms & KDE Plots

These plots visualize the concentration spreads of parameters such as pH, and Dissolved Oxygen across the entire dataset. By observing the overlap of density curves between different pollution categories, we can identify which features have distinct signatures and which are more ambiguous. This provides essential information on whether the data requires normalization or transformation to handle skewed feature distributions before training.

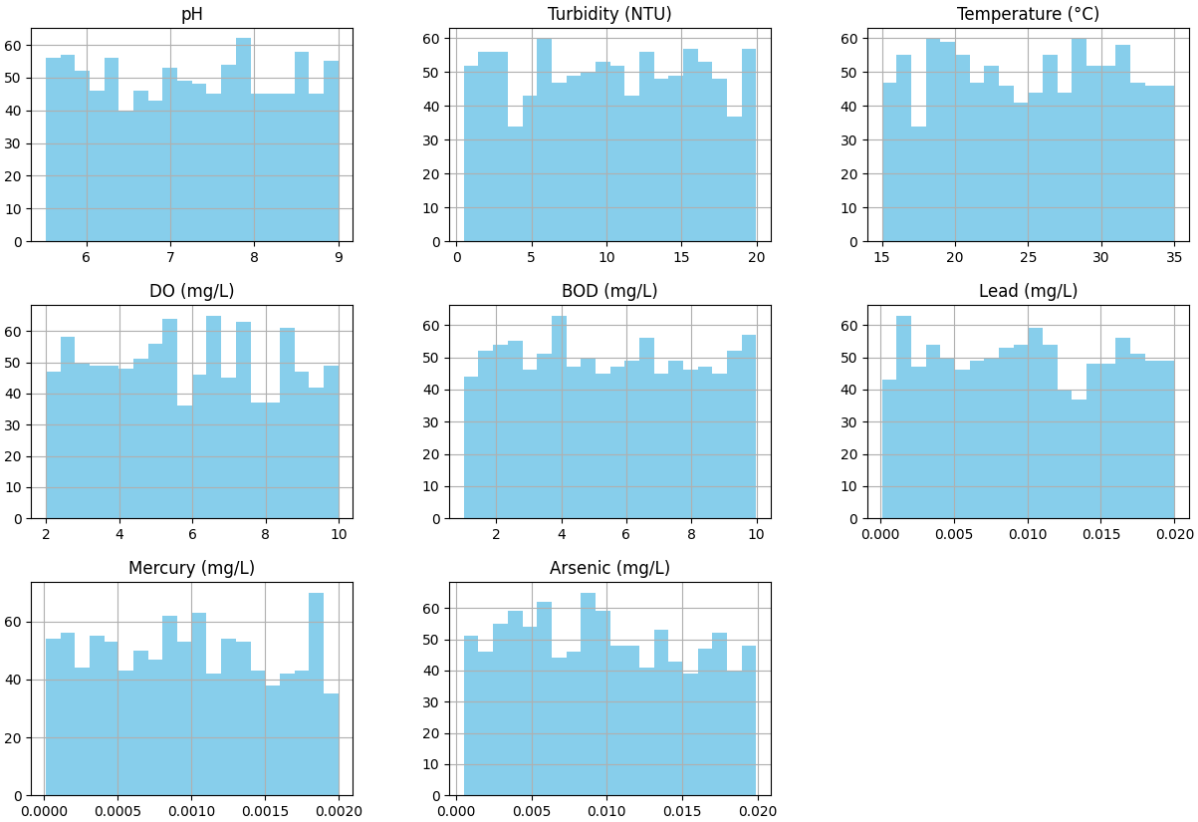


Figure 3.8: Dataset 2 Histograms & KDE Plots

Correlation Heatmap

This visualization explores the linear relationships between various physicochemical properties, such as the link between Mineral Dissolution and Conductivity. High correlation coefficients indicate redundant information that could be streamlined, while low correlations suggest that pollution levels are driven by independent variables. Understanding these inter-dependencies helps in explaining why certain ensemble models may perform better by capturing complex feature interactions.

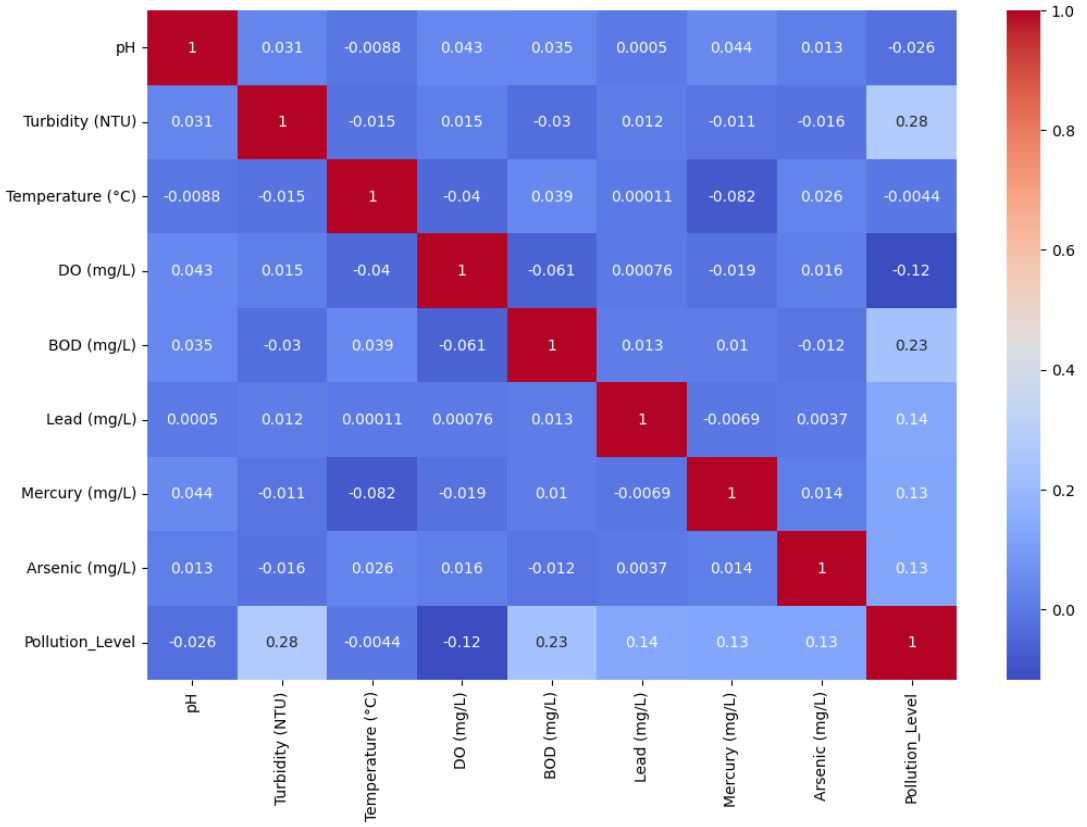


Figure 3.9: Dataset 2 Correlation Heatmap

Boxplots by Event

By grouping physical and chemical features by their assigned pollution tier, boxplots highlight the specific thresholds and outliers that characterize each class. A clear separation in medians for a feature like Turbidity across classes indicates it is a strong discriminative predictor for the model. These plots also help in identifying extreme values that may represent sensor errors or significant, localized contamination events.

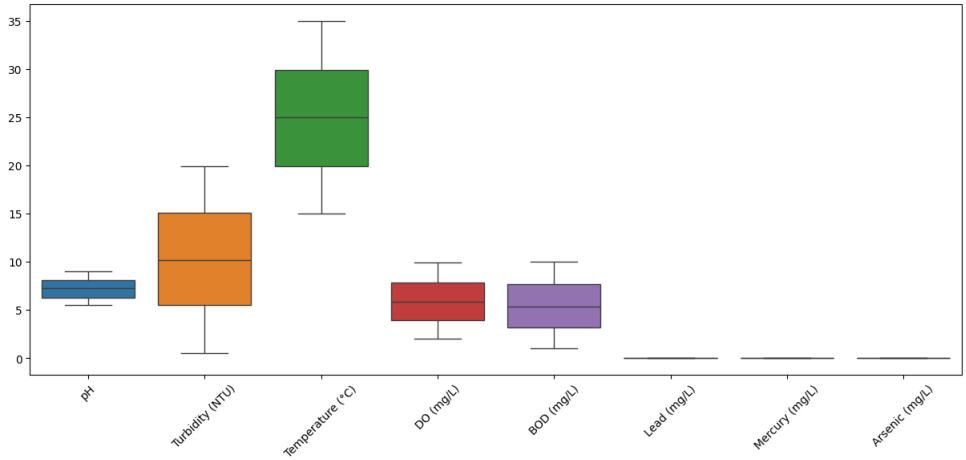


Figure 3.10: Dataset 2 Boxplots by Event

Pair Plots

Pair plots provide a high-dimensional view of the dataset by plotting pairs of features against each other, colored by their respective pollution levels. This allows for the observation of clusters and the degree of overlap between categories, such as "Slightly" versus "Moderately" polluted. Seeing these non-linear boundaries visually justifies the selection of sophisticated algorithms like Gradient Boosting or Neural Networks over simpler linear classifiers.

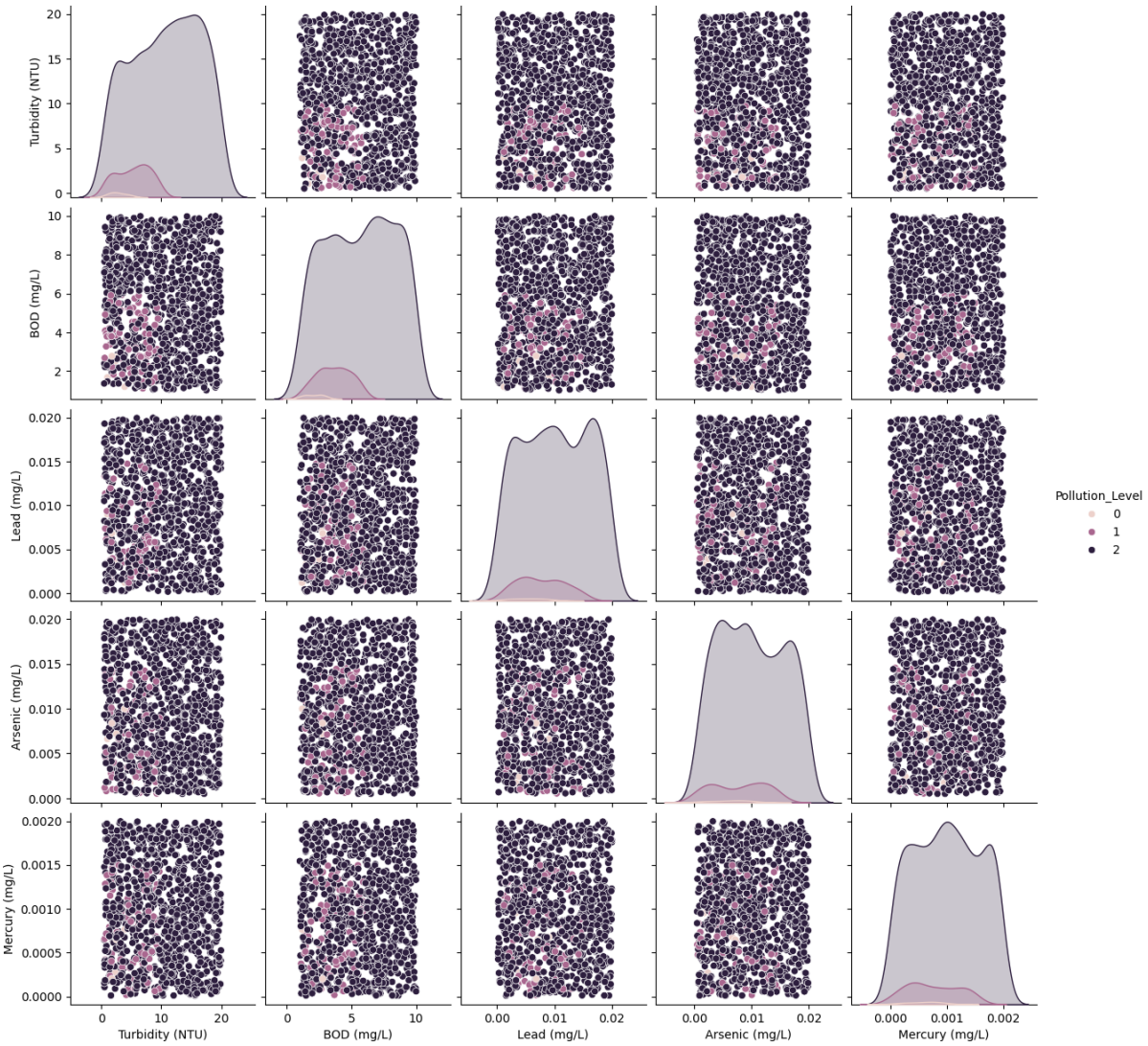


Figure 3.11: Dataset 2 Pair Plots

3.4 Dataset 3: GECCO Industrial IoT Time-Series

The third dataset focuses on high-resolution sensor data from the GECCO (Genetic and Evolutionary Computation Conference) Water Quality challenge,[2] representing a real-

[2]Official challenge page: https://www.spotseven.de/gecco/gecco-challenge/gecco-challenge-2018/

world Industrial IoT (IIoT) monitoring scenario. The GECCO benchmark has been investigated in recent ML studies on multivariate time-series anomaly detection [35].

3.4.1 Dataset Description

This dataset consists of time-series measurements captured by physical sensors deployed in a water distribution network. Unlike the previous chemical-heavy datasets, this data emphasizes the temporal dynamics of water quality.

The dataset captures **real-time sensor streams** from continuous monitoring of physical parameters such as pH, Conductivity, Temperature, and Turbidity. These sensors generate high-frequency data, reflecting the rapid changes that can occur in an industrial or urban water grid due to contamination events or mechanical failures.

Column	Unit	Min (Normal)	Max (Normal)	Description
Tp	°C	0	25	Water temperature
pH	—	6.5	8.5	Acidity/alkalinity of the water
Cond	S/m	0.005	0.05	Electrical conductivity, proxy for dissolved ion concentration
Turb	NTU	0	1	Turbidity, measuring suspended particulate matter
SAC	m^{-1}	0	5	Spectral Absorption Coefficient at 254 nm, indicative of organic matter
PFM	L/h	Flow-dependent	System-specific	Pulse-frequency modulated flow measurement

Table 3.3: GECCO IIoT Sensor Parameters

For the **identification of contamination events**, the target labels are derived from specific "events" where the sensor readings deviate significantly from established baseline behaviors. Detecting these events is critical for early warning systems, as they represent potential threats to the integrity of the water supply in real-time.

3.4.2 Exploratory Data Analysis (EDA)

Water Quality Distribution per Target

The target distribution analysis confirms a severe class imbalance inherent in anomaly detection within sensor networks. With 127,100 instances labeled as "Normal" (Class 0) compared to only 1,900 "Event" instances (Class 1), the minority class represents less than 1.5% of the total data. This quantitative evidence necessitates the use of synthetic resampling in Phase 2 to prevent the model from developing a bias toward the "Normal" operational state.

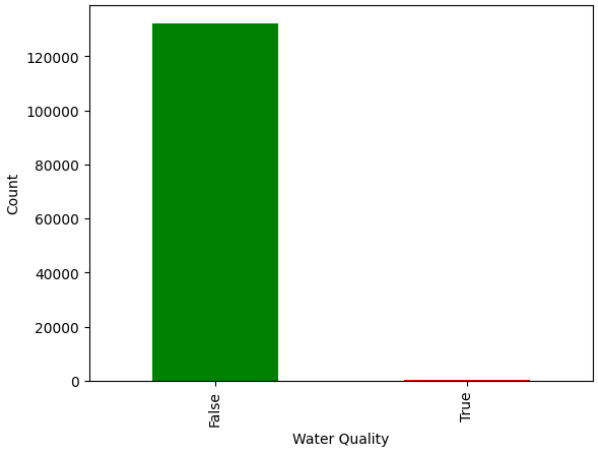


Figure 3.12: Dataset 3 Water Quality Distribution per Target

Histograms & KDE Plots

The density plots for the GECCO sensors reveal varying data spreads across physical parameters like pH, Conductivity, and Turbidity. While some features exhibit a Gaussian-like distribution around a stable operating point, others show multi-modal behavior or heavy tails. These visualizations indicate that the sensor data contains noise and fluctuations that do not always correspond to pollution, requiring the model to distinguish between standard variance and true contamination signatures.

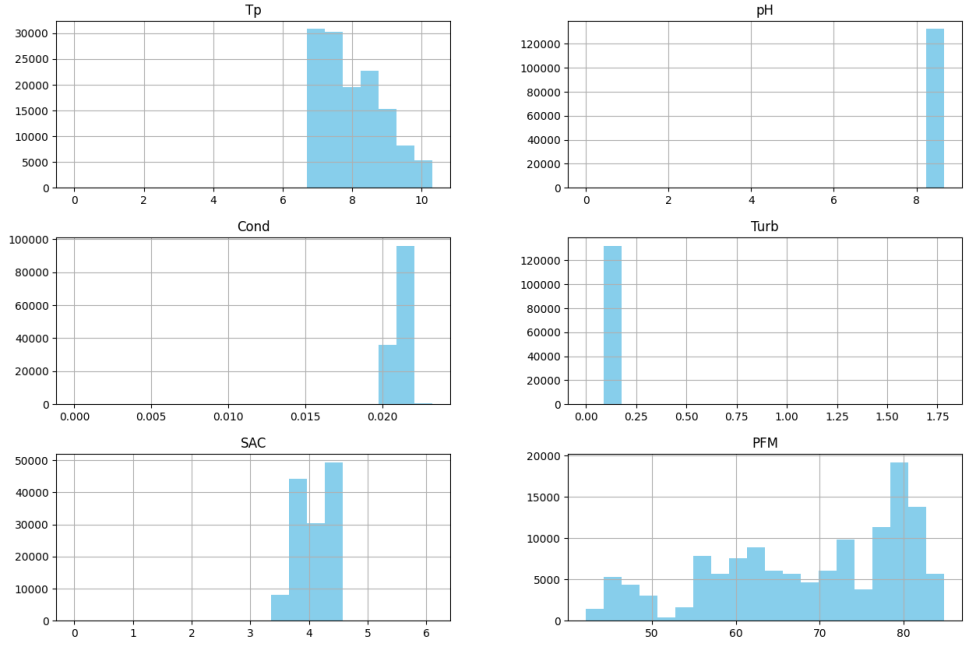


Figure 3.13: Dataset 3 Histograms & KDE Plots

Correlation Heatmap

The correlation matrix highlights the physical relationship between different sensors in the IIoT grid. While certain parameters show moderate positive or negative correlations (e.g., pH vs. Conductivity), many features show near-zero correlation with one another.

This lack of strong linear redundancy suggests that contamination events are complex and multivariate, reinforcing the choice of ensemble methods that can combine weak signals from multiple independent sensors.

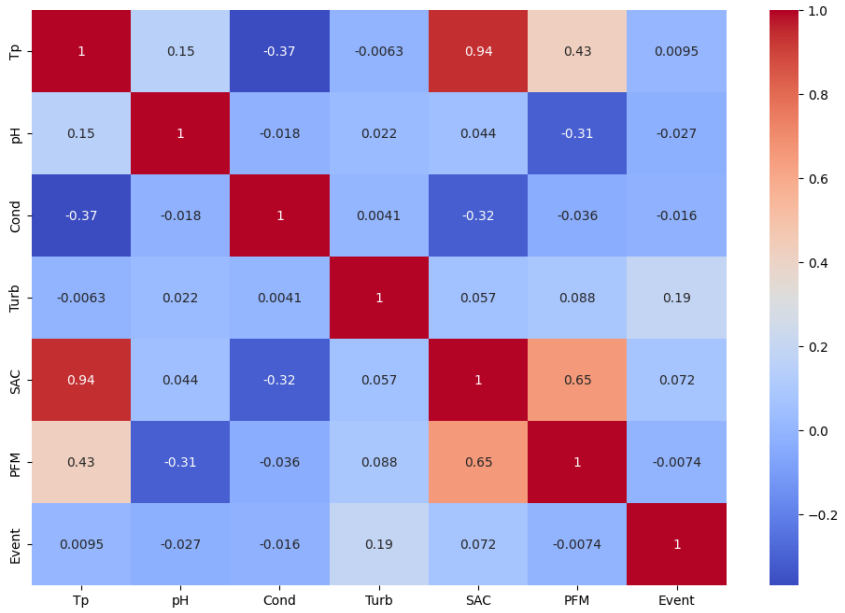


Figure 3.14: Dataset 3 Correlation Heatmap

Boxplots by Event

Grouping the sensor readings by event status reveals distinct "event signatures" where the distribution of the minority class shifts away from the baseline. For several features, the "Event" boxplots show significantly higher variances and multiple extreme outliers compared to the "Normal" state. These visual gaps indicate which sensors are the primary drivers for detection, though the overlap between boxes also highlights the difficulty in setting simple, static alarm thresholds.

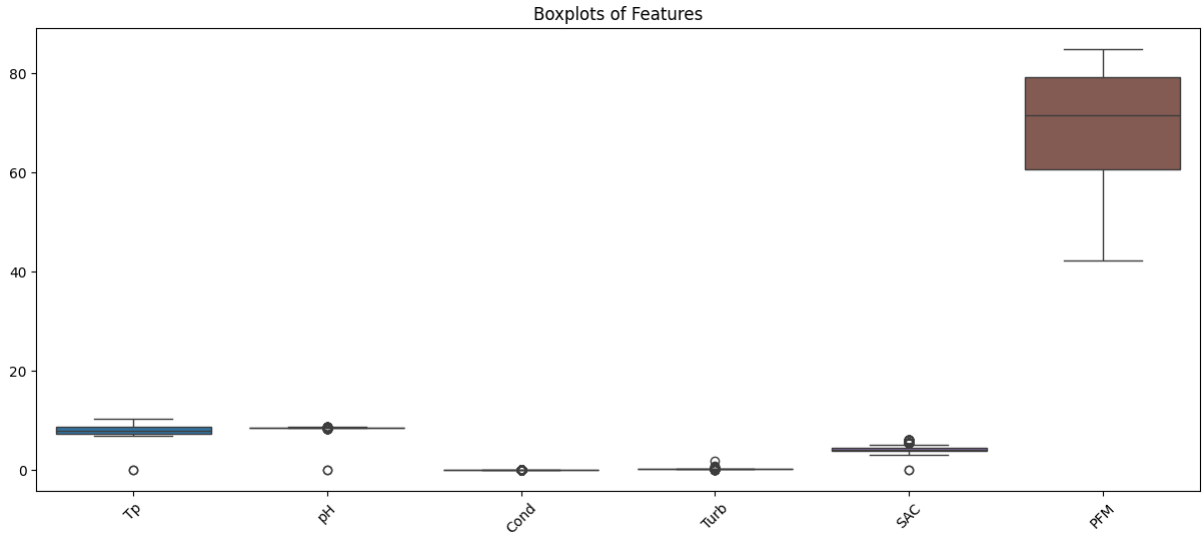


Figure 3.15: Dataset 3 Boxplots by Event

Pair Plots

The pair plot matrix provides a multi-dimensional view of how "Normal" and "Event" states cluster across different sensor combinations. In several projections, the event instances appear as small, scattered clusters within the broader density of normal operation rather than clearly separated groups. This visualization justifies the use of high-capacity non-linear classifiers, such as Extra Trees or Gradient Boosting, which are capable of carving out complex decision boundaries in this high-dimensional feature space.

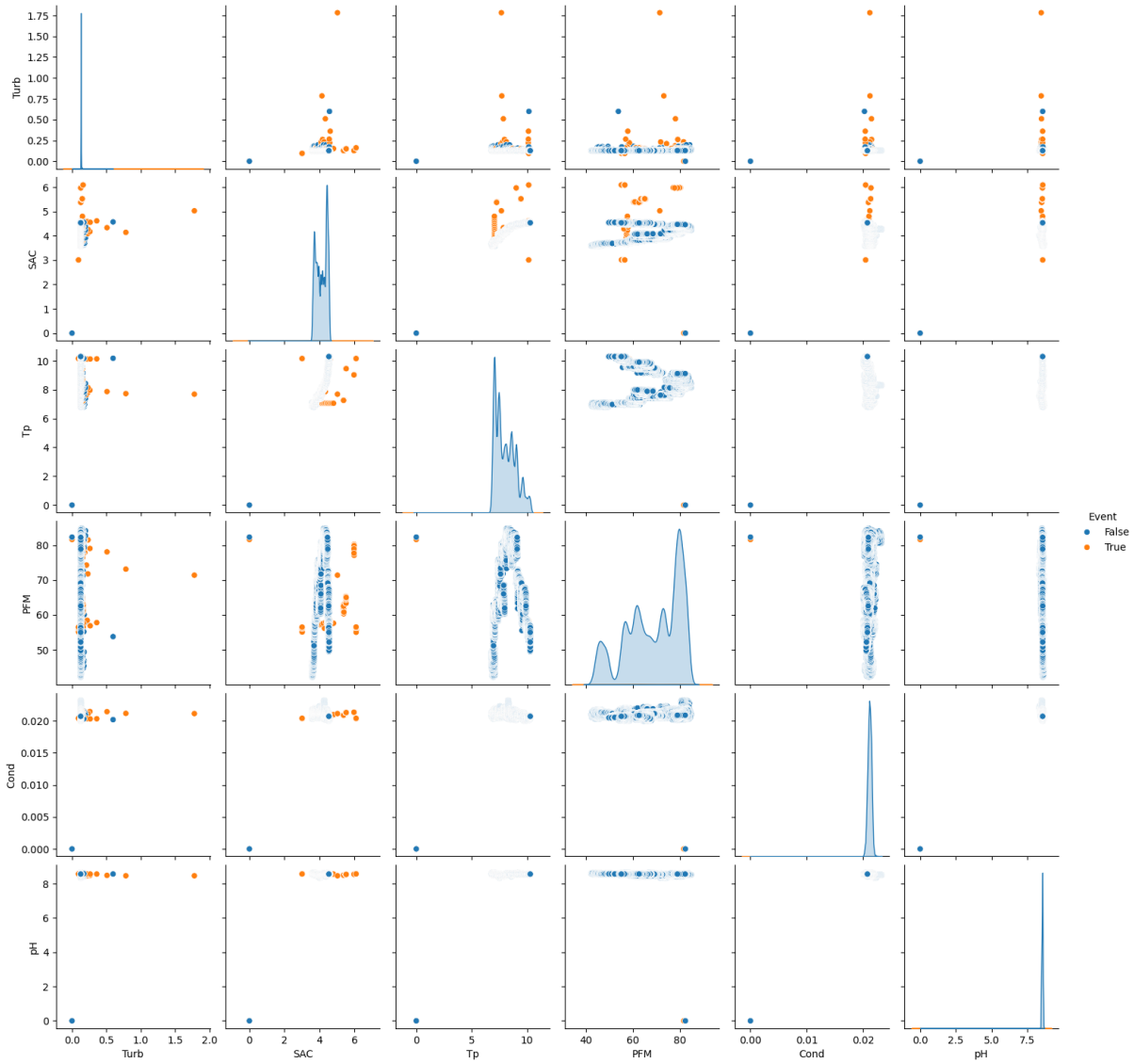


Figure 3.16: Dataset 3 Pair Plots

Chapter 4

Results and Discussion

4.1 Introduction

This chapter presents a comprehensive analysis of the experimental results obtained from applying ensemble machine learning frameworks across the datasets. It presents the detailed results for each dataset, including the feature engineering rationale, phase-by-phase model comparisons, and the selection of the final best-performing model. The chapter also compares the obtained results with state-of-the-art studies, followed by a discussion of the study limitations and potential directions for future research.

4.2 Dataset 1: Binary Water Safety Classification

4.2.1 Dataset Characteristics and the Challenge of Class Imbalance

Dataset 1 comprises 7,996 water samples drawn from an urban environment, each described by 20 physicochemical and biological attributes. The binary target variable `is_safe` partitions the data into 7,084 unsafe samples and 912 safe samples, yielding an imbalance ratio of approximately 7.77:1. This ratio is nonetheless sufficient to introduce systematic bias toward the majority class in standard training procedures if not carefully managed.

The dataset presented no missing values, which simplified the preprocessing pipeline. Nonetheless, because the input features span multiple orders of magnitude and include both near-normal and heavily skewed distributions, careful normalization was required. KNN imputation with five neighbours was applied as a precautionary measure to support generalizability of the pipeline, followed by a Yeo–Johnson Power Transformation to stabilize variance and bring each feature closer to a Gaussian distribution. The data was split 80/20 into training and test sets using stratified sampling, and a ten-fold StratifiedKFold cross-validation scheme was adopted to ensure reliable performance estimates across the full training set.

4.2.2 Feature Engineering: From 20 Raw Features to 40 Discriminative Signals

A central premise of this work is that raw physicochemical measurements, taken in isolation, provide an incomplete picture of water safety. Contaminant toxicity is rarely a linear, additive function of individual concentrations; rather, it often depends on synergistic interactions between multiple substances, non-linear dose–response relationships, and relative concentration ratios. To capture these dynamics, a systematic feature engineering procedure was applied, expanding the original 20-feature space to 40 features.

Pairwise ratio features were computed between the most toxicologically significant pairs of contaminants, reflecting known antagonistic and synergistic relationships in the water quality literature. Polynomial (squared) transformations were applied to the subset of features identified as most important in preliminary ranking analyses, allowing the model to learn accelerating toxicity effects at higher concentrations. Log transformations were applied to features exhibiting right-skewed distributions, improving the numerical conditioning of the gradient-boosted trees. Finally, a set of statistical aggregate features—including inter-feature mean, maximum, and range—was computed to provide a holistic summary of overall contamination burden per sample.

4.2.3 Phase 1 Results: Training Without Imbalance Mitigation

In Phase 1, seven models were trained on the raw imbalanced training set: a tuned CatBoost, a tuned XGBoost, a tuned LightGBM, a tuned Random Forest, Extra Trees, a Stacking Ensemble, and a Soft-Voting Ensemble. Hyperparameter optimization was conducted using Optuna Bayesian optimization for the individual models, and the two ensemble strategies were constructed atop the best-performing base learners.

The results were remarkably strong across the board, demonstrating that the engineered feature space, combined with the inherent regularization properties of gradient-boosted trees, was already highly effective even without addressing the class imbalance. The best model of Phase 1 was the Stacking Ensemble, which achieved an accuracy of 97.69%, an F1-Score of 0.8952, and an AUC-ROC of 0.9847 on the held-out test set. The optimal classification threshold was set at 0.27, determined through threshold optimization on the validation folds to maximize F1 on the minority (safe water) class.

The classification report reveals a clear performance gap between the two classes. The majority class (unsafe, class 0) achieves near-perfect precision (0.98), recall (0.99), and F1-Score (0.99), reflecting the abundance of training examples. The minority class (safe, class 1) achieves precision of 0.92 and recall of 0.87, yielding an F1-Score of 0.90, creditable for a class representing only 11.4% of the training set. The AUC-ROC of 0.9847 and Average Precision of 0.9541 confirm that the model maintains excellent discriminative ability across all threshold settings, not just the operating point.

Table 4.1 shows the Phase 1 training results.

Model	Accuracy	F1-Score	AUC-ROC
CatBoost (tuned)	97.12%	86.47%	99.12%
XGBoost (tuned)	97.62%	88.69%	98.77%
Stacking Ensemble	97.56%	88.63%	98.47%
Soft-Voting Ensemble	97.62%	88.69%	98.35%
LightGBM (tuned)	97.56%	88.50%	98.22%
Random Forest (tuned)	97.31%	86.85%	98.12%
Extra Trees	95.38%	75.97%	97.05%

Table 4.1: Dataset 1, Phase 1: Individual Model Performance Comparison

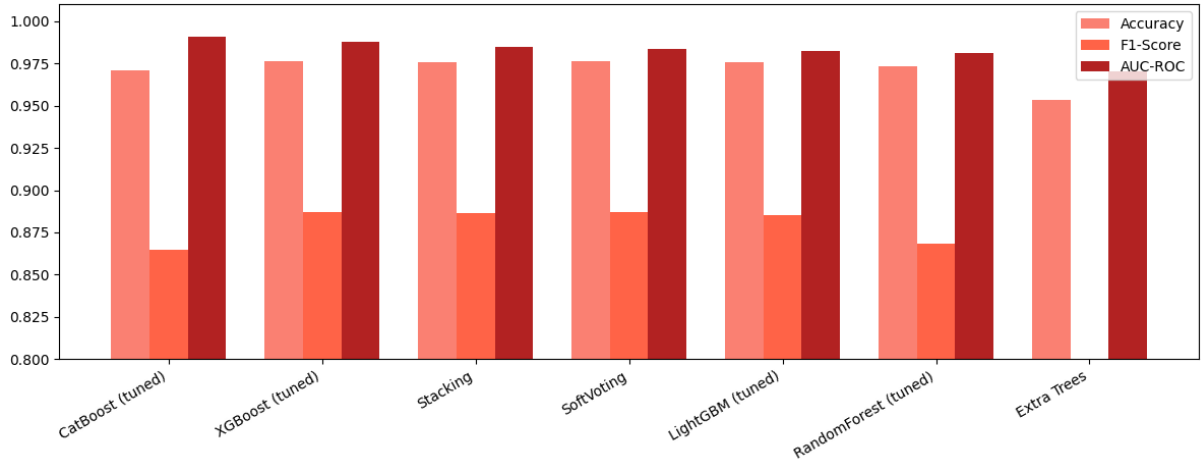


Figure 4.1: Phase 1 Model Comparison Plot

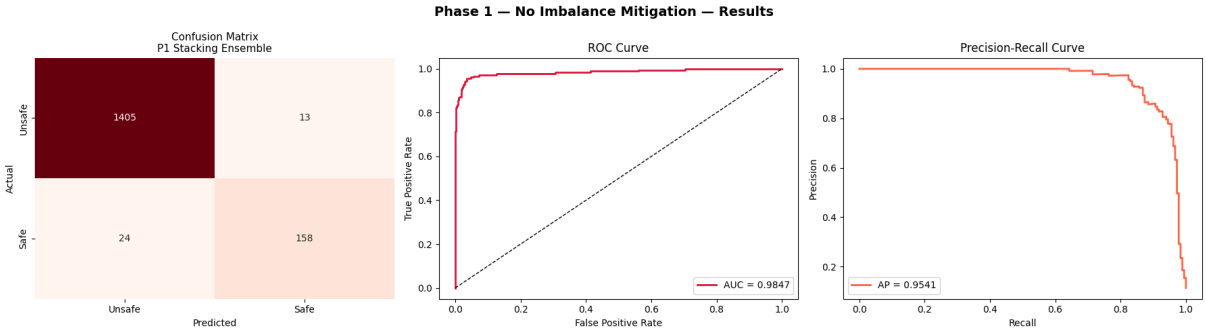


Figure 4.2: Phase 1: Confusion Matrix, ROC Curve, and Precision-Recall Curve

4.2.4 Phase 2 Results: Training With Resampling

Phase 2 introduced SMOTETomek as the selected imbalance mitigation strategy, combining SMOTE oversampling of the minority class with Tomek link removal to clean ambiguous majority-class boundary samples. Applied within the cross-validation loop, this technique balanced the training distribution from a 5,666:730 majority-to-minority

ratio to a near-perfect 5,665:5,665 balance, adding 4,934 synthetic minority samples. The intent was to provide the models with more exposure to the underrepresented safe-water class during training, hypothesizing that this would improve minority-class recall.

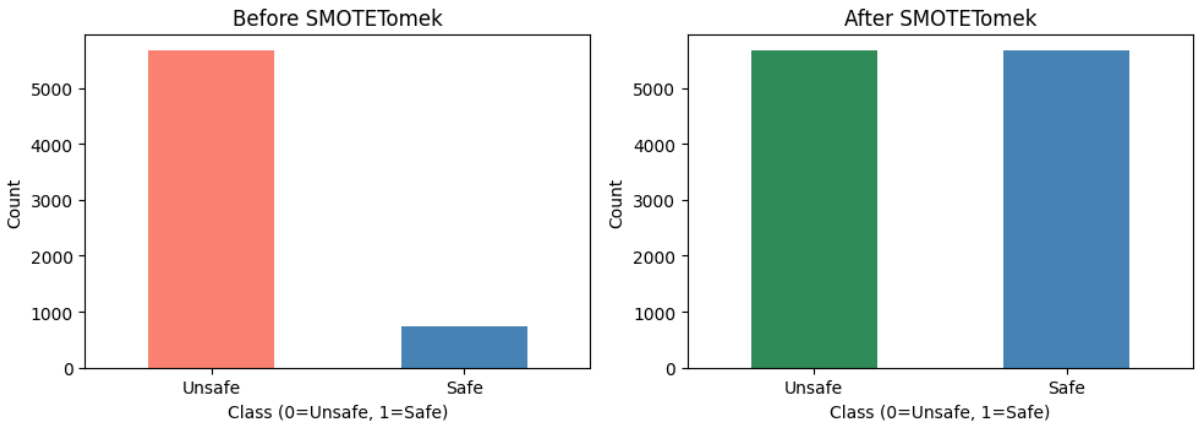


Figure 4.3: Class Distribution Before vs. After SMOTETomek

Contrary to this expectation, Phase 2 produced uniformly weaker results than Phase 1 across all four primary metrics. The Phase 2 best model, a Soft-Voting Ensemble with an optimal threshold of 0.41, achieved an accuracy of 97.12%, an F1-Score of 0.8693, an AUC-ROC of 0.9846, and an Average Precision of 0.9387—declines of 0.56%, 0.0259, 0.0001, and 0.0154, respectively, relative to Phase 1.

Table 4.2 shows the model performance for Phase 2.

Model	**Accuracy**	**F1-Score**	**AUC-ROC**
LightGBM (tuned)	97.25%	87.13%	98.49%
XGBoost (tuned)	97.00%	86.21%	98.73%
Soft-Voting Ensemble	97.00%	85.96%	98.46%
Stacking Ensemble	96.62%	83.73%	98.08%
Random Forest (tuned)	96.56%	83.97%	98.42%
CatBoost (tuned)	96.44%	83.57%	98.65%
Extra Trees	95.94%	79.88%	97.47%

Table 4.2: Dataset 1, Phase 2: Individual Model Performance Comparison

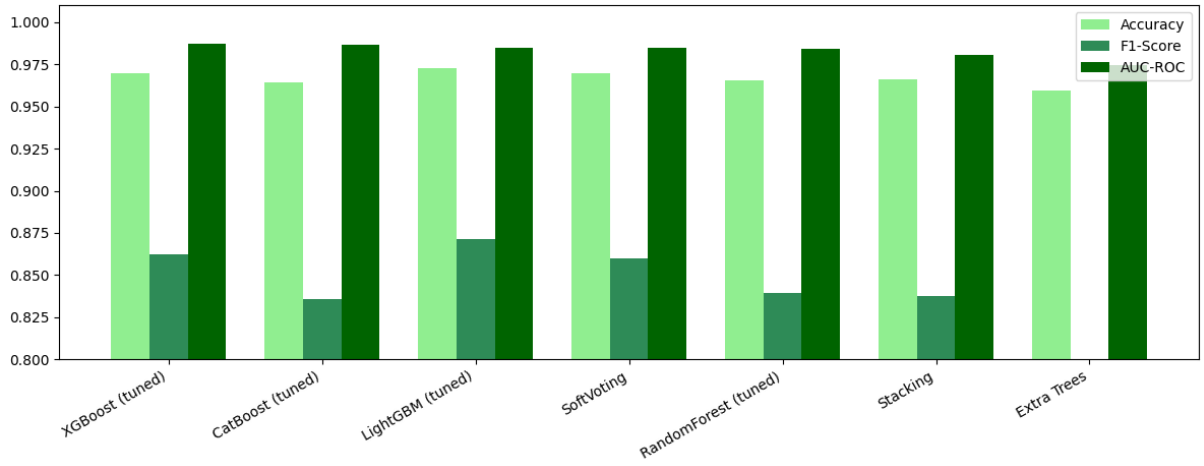


Figure 4.4: Phase 2 Model Comparison Plot

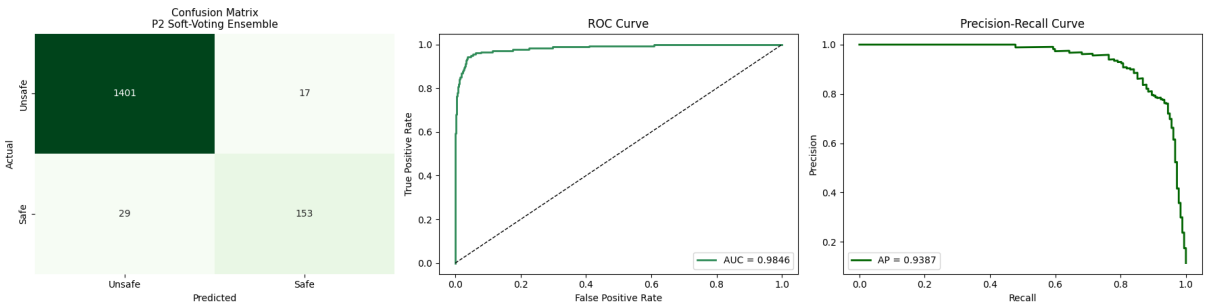


Figure 4.5: Phase 2: Confusion Matrix, ROC Curve, and Precision-Recall Curve

This counterintuitive outcome warrants careful interpretation. The failure of SMOTE-Tomek to improve upon the unmitigated baseline can be attributed to several interacting factors. First, the synthetic minority samples generated by SMOTE are interpolations in feature space between existing minority samples. Given that the engineered feature space contains interaction ratios and polynomial terms, these interpolations may not correspond to physically plausible water compositions, effectively introducing noise rather than genuine signal. Second, gradient-boosted tree ensembles are inherently robust to moderate class imbalance due to their sequential error-correction mechanism, which naturally allocates greater attention to misclassified samples through the boosting weight updates. SMOTETomek therefore provides a benefit that is largely redundant for these model families. Third, the removal of Tomek links near the decision boundary may inadvertently discard informative majority-class samples that help the model learn the contour of the safe-water region.

Table 4.3 compares the best model performance between Phase 1 and Phase 2.

Metric	Phase 1 (No Mitigation)	Phase 2 (SMOTETomek)	Δ
Model	Stacking Ensemble	Soft-Voting Ensemble	—
Accuracy	97.69%	97.12%	−0.56%
F1-Score	0.8952	0.8693	−0.0259
AUC-ROC	0.9847	0.9846	−0.0001
Average Precision	0.9541	0.9387	−0.0154

Table 4.3: Dataset 1: Phase 1 vs. Phase 2 Final Champion Comparison

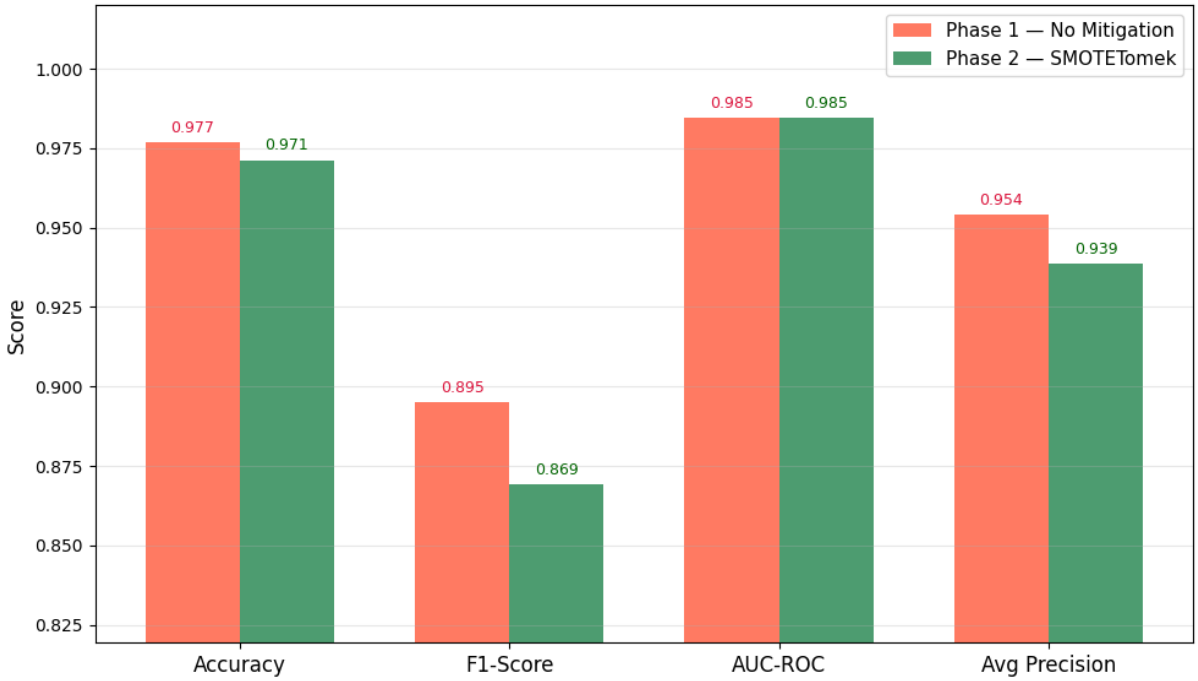


Figure 4.6: Best Model Metrics Comparison Between Phase 1 and Phase 2

Overall, the Phase 1 Stacking Ensemble was selected as the Dataset 1 final model, with its superiority on the minority-class F1-Score serving as the primary selection criterion. This result suggests that for moderate imbalance ratios in the 7–8:1 range, well-engineered features combined with optimized gradient-boosted ensembles can achieve strong minority-class detection without requiring synthetic data augmentation.

4.3 Dataset 2: Multi-Class Pollution Level Classification

4.3.1 Dataset Characteristics and the Extreme Imbalance Challenge

Dataset 2 targets a fundamentally different classification problem: the prediction of water pollution severity across three ordered classes—Level 0 (low pollution), Level 1 (moderate

pollution), and Level 2 (high pollution). The dataset is characterized by ten physicochemical and physical parameters.

The class distribution of Dataset 2 presents a far more severe imbalance problem than Dataset 1. Level 2 (high pollution) accounts for 91.5% of samples, Level 1 (moderate pollution) for 7.9%, and Level 0 (low pollution) for a mere 0.6%, yielding a maximum-to-minimum class ratio of 152.5:1. This extreme skew creates a setting in which a trivially biased classifier that predicts Level 2 for every sample would achieve over 91% accuracy while completely failing on the two minority classes that are arguably the most environmentally critical. Standard accuracy metrics are therefore deeply misleading in this context, and the primary evaluation criteria were shifted to Balanced Accuracy and F1 Macro.

The small dataset size further exacerbated the difficulty. With only approximately 6 training samples belonging to Level 0, even sophisticated resampling cannot fully compensate for the paucity of real minority-class information. A five-fold StratifiedKFold was adopted to ensure that each validation fold contained at least one representative of each class.

4.3.2 Feature Engineering: Constructing Composite Pollution Indices

The eight raw physicochemical features were expanded to 39 through a domain-informed engineering procedure. Polynomial features were derived for the top-correlated variables and pairwise interaction ratios were computed for the four most predictive features. Three composite domain indices were constructed: a Heavy Metal Composite Index (the normalized sum of lead, mercury, and arsenic concentrations), an Organic Load Index (combining BOD and related oxygen-depletion indicators), and a Physical Condition Index (aggregating turbidity and temperature). These indices embed domain knowledge about how pollutant mixtures compound to shift water quality across severity levels. Heavily skewed features (skewness > 1) were log-transformed to reduce the leverage of extreme contamination events on gradient estimates during model training.

4.3.3 Phase 1 Results: With Imbalance Handling (SMOTE-ENN)

In Dataset 2, the phases are labelled according to the imbalance handling strategy rather than its presence or absence. Phase 1 applied SMOTE-ENN—a combined oversampling and cleaning strategy—together with class-weight balancing and calibrated probability outputs, within the cross-validation loop. Fourteen models were evaluated at baseline.

Table 4.4 summarizes the Phase 1 training results.

Model	Accuracy	Bal. Acc.	F1 Macro	F1 Weighted	F1 Macro Std
CatBoost	0.94625	0.862158	0.756435	0.951707	0.18451
Gradient Boosting	0.95375	0.850512	0.735849	0.957221	0.159365
XGBoost	0.94500	0.821887	0.721766	0.949804	0.164838
LightGBM	0.93875	0.839039	0.713738	0.945110	0.156086
Extra Trees	0.92125	0.798665	0.710356	0.929714	0.169619
Random Forest	0.93375	0.816864	0.694938	0.940118	0.167342
Decision Tree	0.92625	0.783641	0.683760	0.933535	0.155613
MLP (Neural Net)	0.89250	0.775074	0.640545	0.909565	0.171920
SVM	0.89750	0.785960	0.630471	0.911418	0.159065
Naive Bayes	0.85750	0.776536	0.615562	0.883552	0.165521

Table 4.4: Dataset 2, Phase 1: Model Performance Summary (SMOTE-ENN)

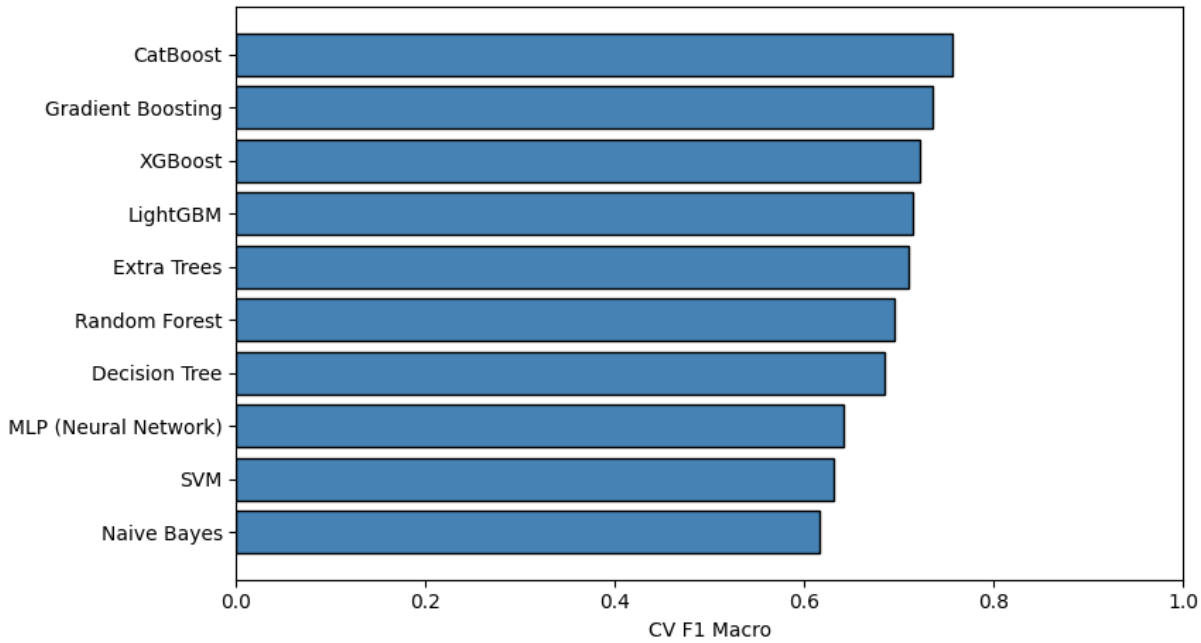


Figure 4.7: Dataset 2, Phase 1: Model Performance Summary Chart

CatBoost emerged as the strongest baseline model with a CV F1 Macro of 0.7564 and Balanced Accuracy of 0.8622. Following Optuna hyperparameter optimization and construction of a stacking ensemble, the tuned CatBoost achieved a test-set F1 Macro of 0.7151, Balanced Accuracy of 0.9266, MCC of 0.7552, and AUC-ROC of 0.9939. The high AUC-ROC indicates excellent probabilistic discrimination across all class pairs, while the F1 Macro reflects the ongoing difficulty of achieving strong recall on Level 0.

The classification report for the Phase 1 CatBoost is revealing. Level 2 achieves near-perfect precision (0.99) and recall (0.97). Level 1 achieves precision of 0.72 and recall of 0.81, indicating that the model can reliably identify moderately polluted water when the signal is present. Level 0 achieves precision of 0.25 and perfect recall of 1.00, meaning the model correctly identifies the single Level 0 test sample but at the cost of generating false positives. The stacking ensemble achieves slightly lower F1 Macro (0.6966) but with a

more conservative prediction pattern for Level 1 (precision 1.00, recall 0.44), illustrating the precision–recall trade-off inherent in extreme imbalance settings.

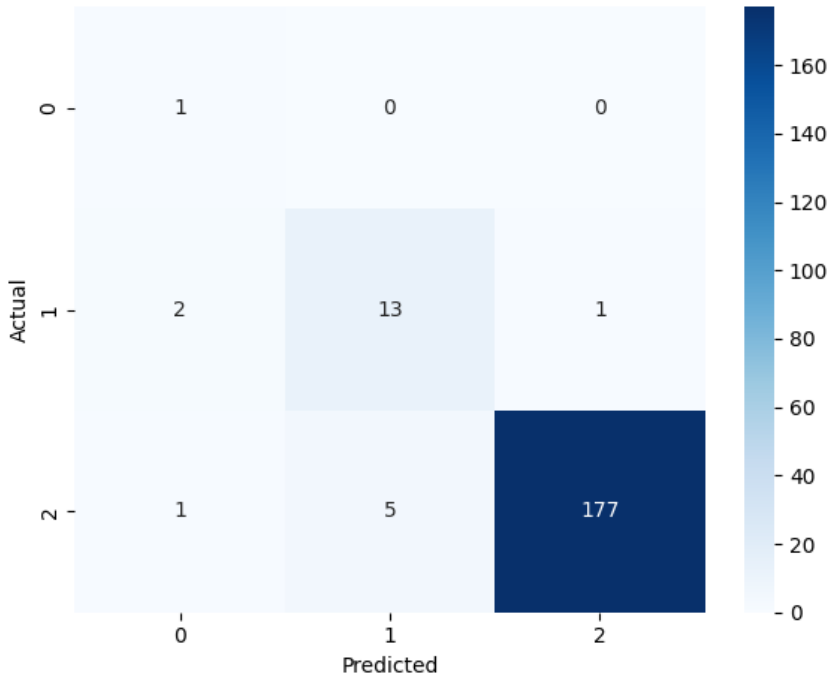


Figure 4.8: Test Set Results: CatBoost Model in Phase 1

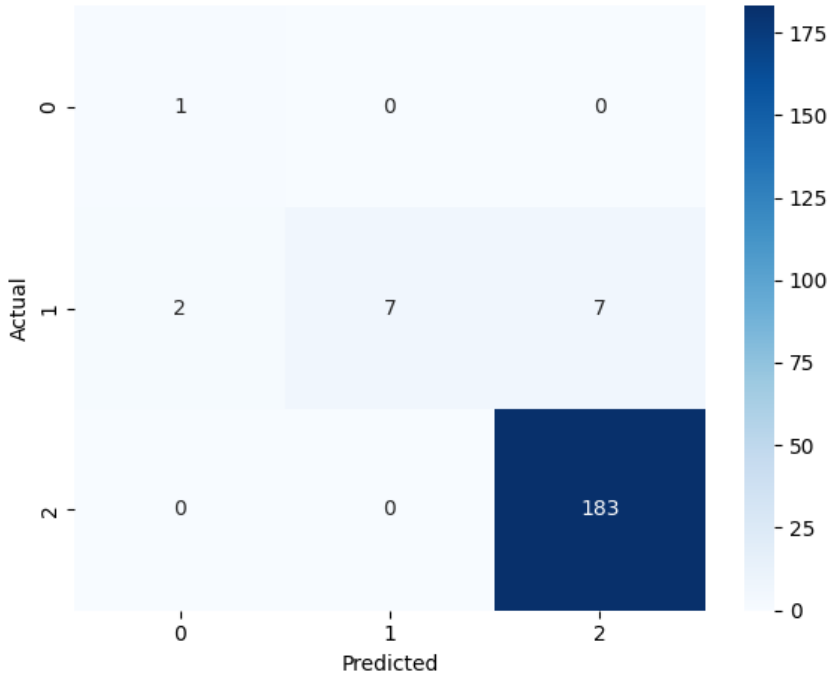


Figure 4.9: Test Set Results: Stacking Ensemble in Phase 1

4.3.4 Phase 2 Results: Without Resampling (Cost-Sensitive Native Handling)

Phase 2 eschewed any synthetic data generation entirely and relied instead on the native cost-sensitive mechanisms built into each model and calibrated probabilities. This approach tests whether the internal weighting mechanisms of modern boosting algorithms are sufficient to handle extreme multi-class imbalance without external resampling.

Table 4.5 summarizes the Phase 2 training results.

Model	Accuracy	Bal. Acc.	F1 Macro	F1 Weighted	F1 Macro Std
CatBoost	0.99000	0.906973	0.875751	0.989175	0.15777
LightGBM	0.98625	0.862528	0.862308	0.984078	0.17048
XGBoost	0.98500	0.790079	0.794273	0.981572	0.189545
Gradient Boosting	0.98375	0.796081	0.792353	0.980529	0.192083
Random Forest	0.97500	0.753351	0.761274	0.971244	0.178642
Decision Tree	0.96750	0.819807	0.753084	0.967758	0.17404
SVM	0.89375	0.778797	0.632691	0.908538	0.160331
Naive Bayes	0.87250	0.809320	0.623638	0.895341	0.152928
Logistic Regression	0.87625	0.845164	0.623450	0.898998	0.143729
MLP (Neural Net)	0.93500	0.605282	0.606279	0.925412	0.146833

Table 4.5: Dataset 2, Phase 2: Model Performance Summary (No Resampling)

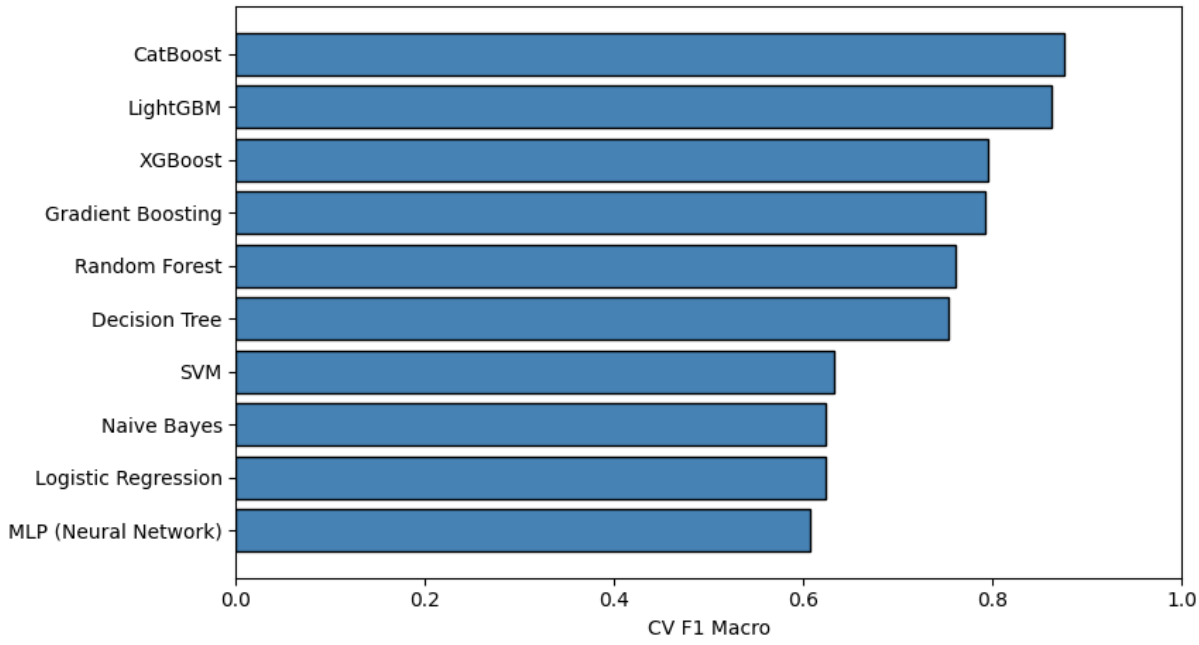


Figure 4.10: Dataset 2, Phase 2: Model Performance Summary Chart

The results of Phase 2 were decisively superior to Phase 1 across all major metrics. CatBoost in Phase 2 achieved a CV F1 Macro of 0.8758, a delta of +0.1193 over Phase 1, and a Balanced Accuracy of 0.9070. On the test set, Phase 2 CatBoost achieved an accuracy of 99.00%, F1 Macro of 0.8658, Balanced Accuracy of 0.9583, MCC of 0.9352,

and AUC-ROC of 0.9991—the strongest results achieved by any single model across any configuration in Dataset 2.

The classification report for Phase 2 CatBoost demonstrates marked improvements. Level 1 now achieves precision of 1.00 and recall of 0.88 (F1 = 0.93), compared to 0.72 precision and 0.81 recall in Phase 1. Level 0 achieves precision of 0.50 and recall of 1.00 (F1 = 0.67), a modest precision improvement over Phase 1's 0.25. These gains suggest that the synthetic samples introduced by SMOTE-ENN in Phase 1 were in fact interfering with the natural decision boundaries learned by CatBoost's ordered boosting mechanism, rather than helping it distinguish between pollution severity levels.

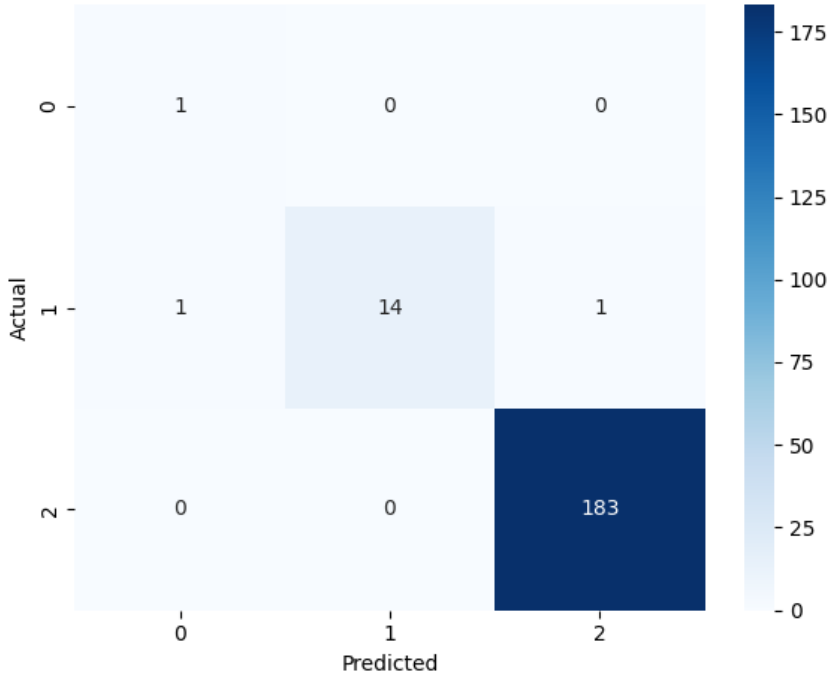


Figure 4.11: Test Set Results: CatBoost Model in Phase 2

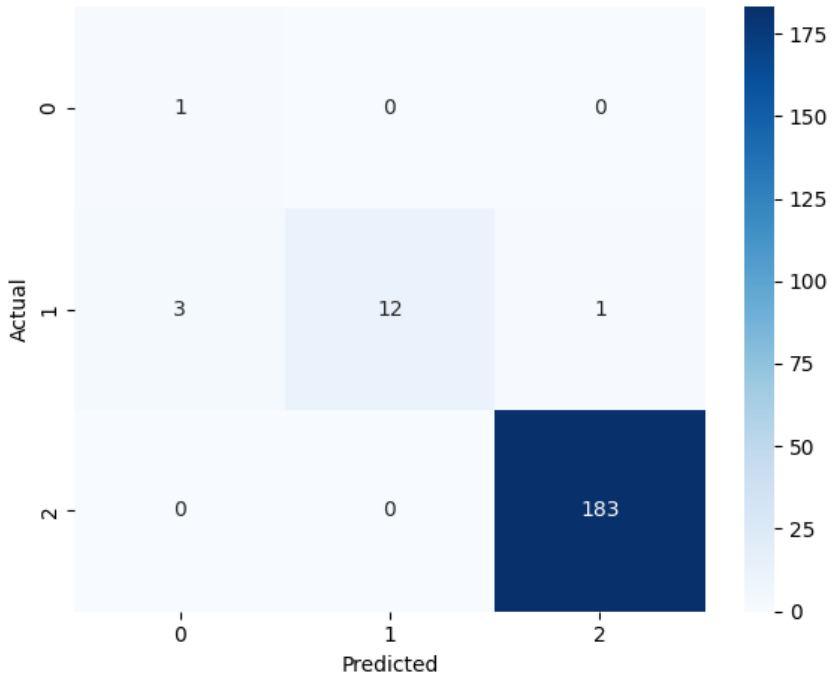


Figure 4.12: Test Set Results: Stacking Ensemble in Phase 2

Table 4.6 compares CatBoost test-set performance across phases.

Metric	**Phase 1 (SMOTE-ENN)**	**Phase 2 (No Resampling)**	**Δ**
Accuracy	0.955	0.990	+0.035
Balanced Accuracy	0.9266	0.9583	+0.0317
F1 Macro	0.7151	0.8658	+0.1507
MCC	0.7552	0.9352	+0.18
AUC-ROC (OvR)	0.9939	0.9991	+0.0052

Table 4.6: Dataset 2: Phase 1 vs. Phase 2 — CatBoost Test-Set Performance

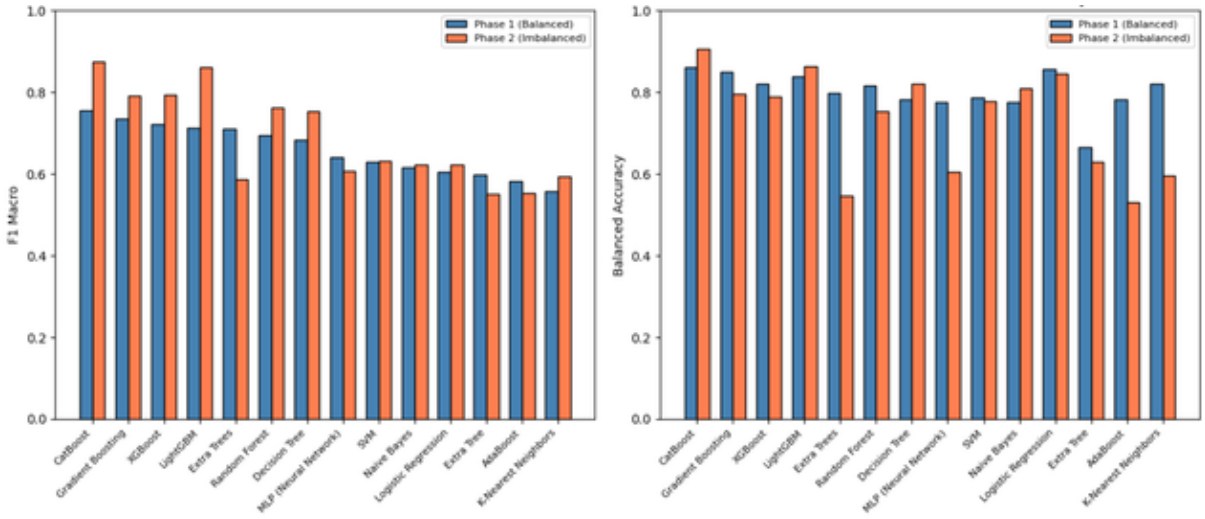


Figure 4.13: F1 Macro and Balanced Accuracy Comparison Between Phase 1 and Phase 2

Across all models evaluated, Phase 2 outperformed Phase 1 in almost all cases by F1 Macro, with the two exceptions being Extra Trees and MLP. Those models have architectures that benefit from resampling because they lack built-in cost-sensitive mechanisms. The average improvement across all models was $\Delta F1 = +0.0260$ in favour of Phase 2, leading to the study's verdict: for this dataset, models are robust to extreme class imbalance without explicit resampling, provided they incorporate internal class weighting mechanisms. This finding has important practical implications, as it suggests that SMOTE-style augmentation on very small datasets with extreme class ratios can paradoxically harm model performance by introducing synthetic boundary ambiguity.

4.4 Dataset 3: Real-Time Anomaly Detection in Drinking Water (GECCO IIoT)

4.4.1 Dataset Characteristics: Large-Scale Temporal Imbalance

Dataset 3 represents a real-world Industrial Internet of Things (IIoT) monitoring scenario. The binary target variable `Event` identifies drinking water contamination events, of which only 628 occur, a positive event rate of approximately 0.29%, corresponding to a class imbalance ratio of 347.5:1.

The contamination events are not randomly distributed across time; they cluster into discrete episodes during which sensor readings deviate from their normal baseline patterns. A model that fails to capture the temporal context of these deviations will produce high accuracy through majority-class dominance while missing almost all true contamination events. This motivates the primary feature engineering strategy for Dataset 3: temporal signal characterization.

4.4.2 Feature Engineering: Temporal Signal Characterization

The six raw sensor readings were transformed into 76 features through a comprehensive temporal feature engineering procedure. Rolling statistics—mean, standard deviation, minimum, and maximum—were computed over windows of 5 and 15 consecutive samples for each sensor, capturing both short-term fluctuations and medium-term trends in the signal. Rate-of-change features (first-order differences per sensor) quantified the velocity of change between consecutive measurements, which is a critical indicator of sudden contamination events. Interaction terms between physically related sensor pairs were computed: turbidity and SAC (both related to particulate and organic matter in water), and conductivity and pH (both influenced by ionic composition), reflecting known physicochemical coupling in drinking water systems. Anomaly indicator features were constructed using z-score thresholding (flags for observations exceeding ± 3 standard deviations from the rolling mean), providing explicit binary signals of sensor departure from established norms.

The resulting 76-feature dataset encodes not only the instantaneous state of each sensor but also the dynamic trajectory of the monitored water system, enabling models to distinguish between sustained normal variation and abrupt contamination-related deviations.

4.4.3 Phase 1 Results: Strong Baseline Without Imbalance Mitigation

Phase 1 evaluated models on the raw imbalanced dataset. The results were exceptional, with the top five models all exceeding an F1-Score of 0.91.

Table 4.7 shows all model results for Phase 1.

Model	F1	AUC-ROC	PR-AUC	Precision	Recall	Accuracy
Extra Trees	0.9514	0.9976	0.9627	0.9872	0.9183	0.9997
CatBoost	0.9368	0.9971	0.9504	0.9910	0.8884	0.9997
Random Forest	0.9291	0.9955	0.9551	0.9886	0.8765	0.9996
XGBoost	0.9276	0.9970	0.9494	0.9778	0.8825	0.9996
MLP	0.8469	0.9895	0.8697	0.9293	0.7847	0.9992
Gradient Boosting	0.8461	0.9753	0.7251	0.8808	0.8147	0.9991
Logistic Regression	0.7474	0.9663	0.7503	0.9201	0.6335	0.9988

Table 4.7: Dataset 3, Phase 1: Model Performance on Raw Imbalanced Data

Extra Trees emerged as the clear leader with an F1-Score of 0.9514, AUC-ROC of 0.9976, and PR-AUC of 0.9627—the last being a particularly important metric for severely imbalanced binary classification, as it measures performance in the precision–recall space where the positive class is rare. CatBoost followed with an F1-Score of 0.9368, and Random Forest with 0.9291.

The dominance of Extra Trees in this context is notable and reflects the structure of the GECCO dataset. Contamination events, while rare, tend to produce distinctive, well-separated sensor signatures once the temporal features are computed. Extra Trees' randomized threshold selection at each split, combined with its full-depth tree growth, makes it particularly effective at carving out compact, high-purity regions in high-dimensional

feature spaces—precisely the kind of localized decision regions that correspond to brief, intense contamination episodes.

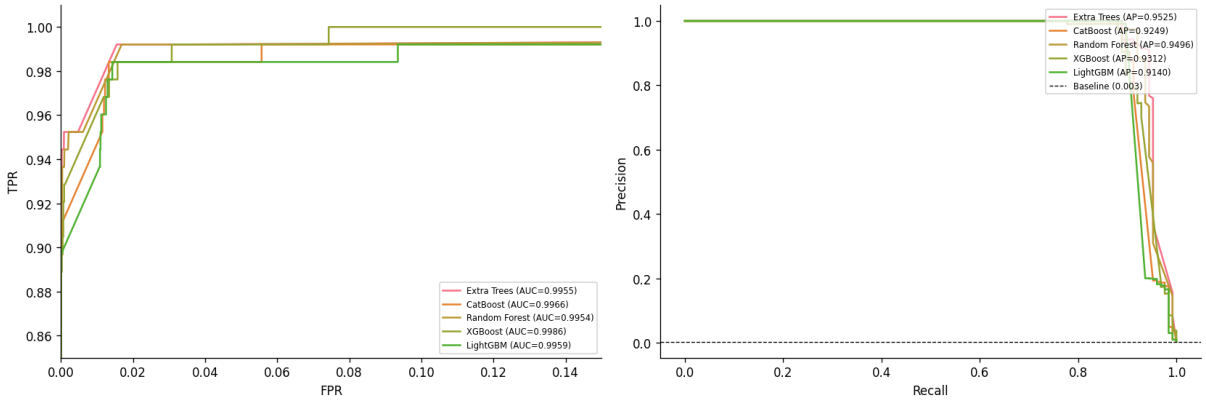


Figure 4.14: Phase 1: ROC and Precision-Recall Curves

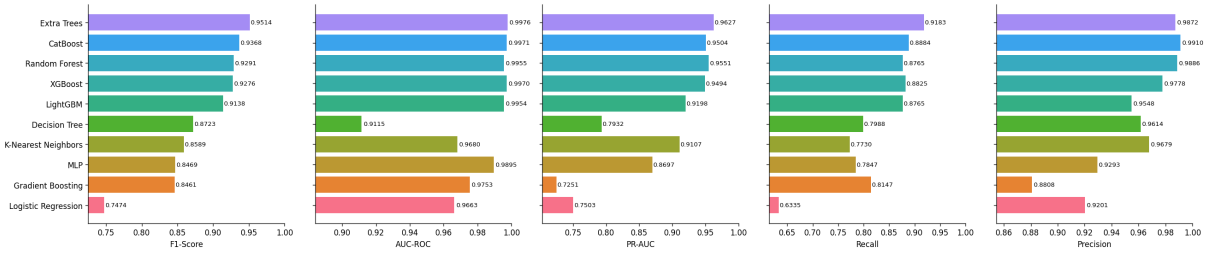


Figure 4.15: Phase 1: Comparison of Each Metric Across All Models

On the test set, the Phase 1 Extra Trees model achieved a precision of 0.9911 and recall of 0.8810 for the positive class, yielding an F1-Score of 0.9328 and Balanced Accuracy of 0.9405. These results confirm that the rich temporal feature engineering effectively transformed the raw time-series imbalance problem into a well-conditioned tabular classification problem in which modern ensemble methods can achieve near-state-of-the-art performance without any resampling.

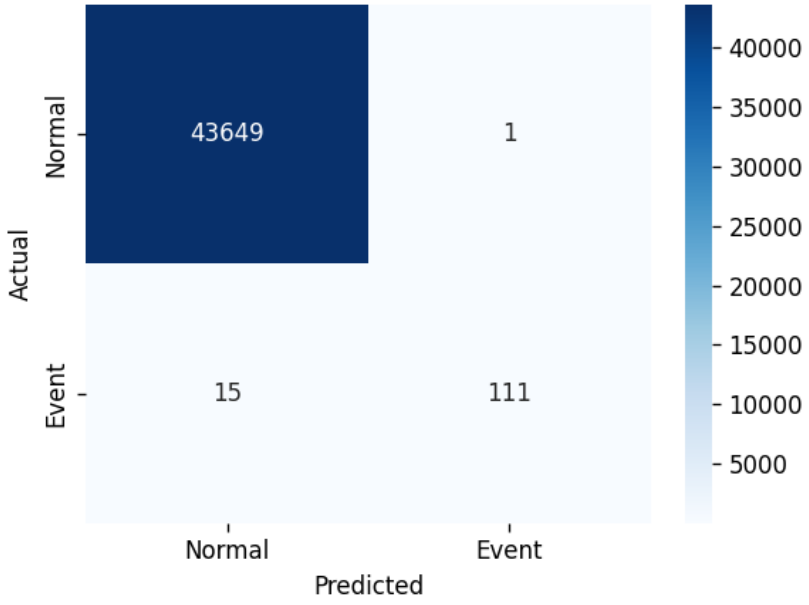


Figure 4.16: Phase 1: Confusion Matrix for Best Model (Extra Trees) on Test Set

4.4.4 Phase 2 Results: Imbalance Mitigation and Optuna Tuning

Phase 2 evaluated four resampling strategies (SMOTE, BorderlineSMOTE, SMOTETomek, and SMOTEENN) using Random Forest as a probe model to identify the optimal strategy before applying it to all models. BorderlineSMOTE achieved the highest CV F1-Score of 0.9182, outperforming SMOTE (0.3563), SMOTETomek (0.3563), and SMOTEENN (0.9080). The superiority of BorderlineSMOTE over standard SMOTE is mechanistically coherent: by focusing synthetic sample generation specifically on the borderline minority samples that are most difficult to classify, it avoids populating the interior of the minority-class cluster with redundant synthetic instances and instead reinforces the decision boundary precisely where it matters most.

Table 4.8 shows all model results for Phase 2.

Model	F1	AUC-ROC	PR-AUC	Precision	Recall	Accuracy
Extra Trees	0.9350	0.9994	0.9592	0.9524	0.9183	0.9996
Random Forest	0.9253	0.9980	0.9484	0.9517	0.9004	0.9996
CatBoost	0.9120	0.9959	0.9434	0.8963	0.9283	0.9995
Gradient Boosting	0.8931	0.9929	0.8671	0.8641	0.9243	0.9994
MLP	0.8761	0.9871	0.8834	0.8558	0.8984	0.9993
Logistic Regression	0.4373	0.9581	0.6831	0.2929	0.8665	0.9936
XGBoost	0.3645	0.9963	0.9459	0.2242	0.9761	0.9902

Table 4.8: Dataset 3, Phase 2: Model Performance with BorderlineSMOTE

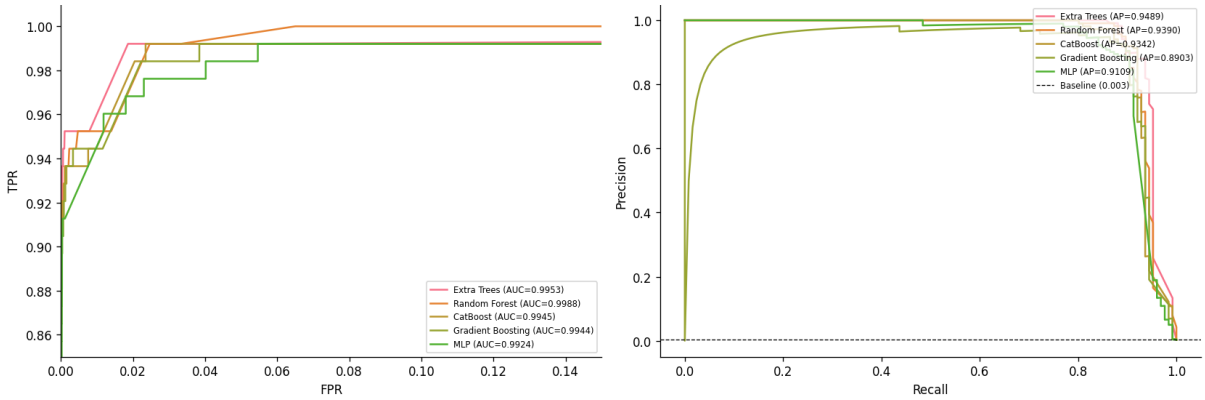


Figure 4.17: Phase 2: ROC and Precision-Recall Curves

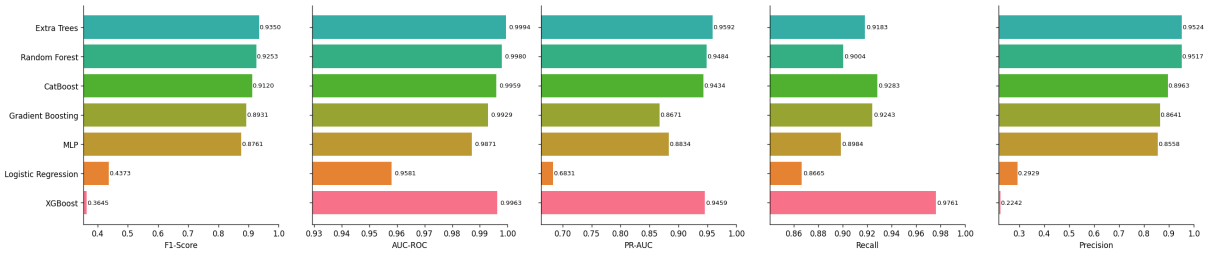


Figure 4.18: Phase 2: Comparison of Each Metric Across All Models

Phase 2 models trained with BorderlineSMOTE generally underperformed Phase 1 equivalents on the F1-Score metric, consistent with the pattern observed in Dataset 1. The Phase 2 Extra Trees model achieved an F1-Score of 0.9350 compared to 0.9514 in Phase 1, while Gradient Boosting showed the opposite trend, improving from 0.8461 (Phase 1) to 0.8931 (Phase 2). This divergence reflects the fundamental difference between tree ensembles with built-in randomization (which handle imbalance implicitly) and those that aggregate deterministic weak learners (which benefit more from external oversampling).

Following Phase 2 cross-validation, Optuna Bayesian hyperparameter optimization was applied to the three best-performing boosting models. Tuned LightGBM emerged as the strongest Phase 2+ model with an F1-Score of 0.9482, AUC-ROC of 0.9968, and PR-AUC of 0.9564—the only model to surpass Phase 1 Extra Trees on the F1-Score metric. The tuned XGBoost and CatBoost followed with F1-Scores of 0.9309 and 0.9220, respectively.

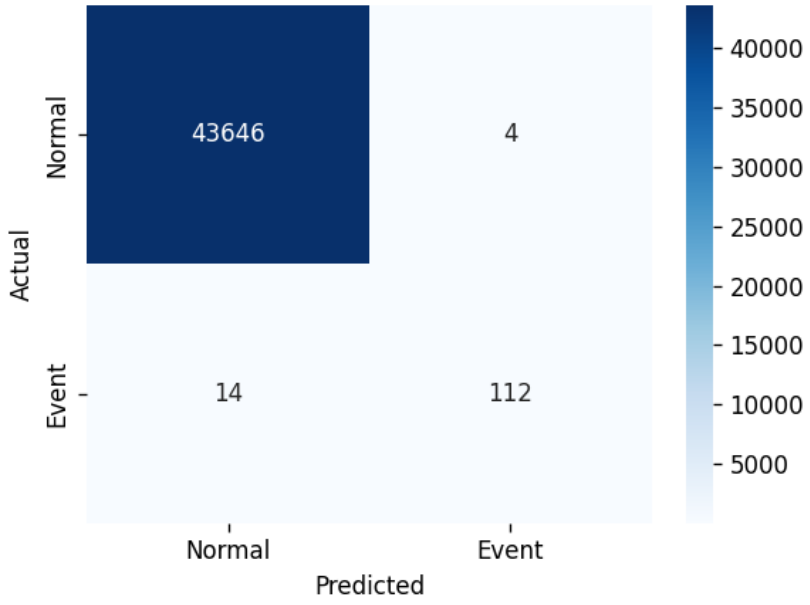


Figure 4.19: Phase 2: Confusion Matrix for Best Model on Test Set

Table 4.9 summarizes the best trained models across both phases.

Phase / Model	**F1-Score**	**AUC-ROC**	**PR-AUC**
Phase 1 — Extra Trees	0.9514	0.9976	0.9627
Phase 2+ — LightGBM (Tuned)	0.9482	0.9968	0.9564
Soft-Voting Ensemble	0.9295	0.9986	0.9467
Phase 2 — Extra Trees (BSMOTE)	0.9350	0.9994	0.9592
Phase 1 — CatBoost	0.9368	0.9971	0.9504

Table 4.9: Dataset 3: Final Cross-Configuration Comparison (Selected Models)

4.5 Comparison with State-of-the-Art Literature

4.5.1 Overview of the Comparison Framework

The latest state-of-the-art studies were identified as directly comparable benchmarks, each employing the same or highly similar datasets. Patel et al. (2022) and Fatima et al. investigated binary water safety classification on a Kaggle water quality dataset (corresponding to Dataset 1). Chen et al. (2024) and Elmotawakkil et al. (2025) addressed multi-class groundwater quality prediction problems comparable to Dataset 2. Wang et al. (2025) proposed a frequency-domain Transformer with Mixture-of-Experts gating for anomaly detection on the GECCO 2019 dataset (Dataset 3). The comparison below evaluates the present study against each of these works on the metrics they report, noting differences in dataset version, sample size, and experimental protocol where relevant.

4.5.2 Dataset 1: Comparison Against Patel et al. (2022) and Fatima et al.

Patel et al. (2022) applied a Random Forest with SMOTE rebalancing and LIME explainability to a version of the Kaggle water potability dataset containing 3,276 samples and 9 features, reporting an accuracy of 81%, precision of 0.83, recall of 0.80, and F1-Score of 0.82. The present study applied the same general model family (ensemble tree methods) to a substantially larger and richer version of the dataset (7,996 samples, 20 raw features, 40 engineered features), achieving an accuracy of 97.69%, precision of 0.92, recall of 0.87, and F1-Score of 0.8952 for the minority (safe water) class. The accuracy improvement of approximately 17 percentage points and the F1-Score improvement of 0.08 reflect both the richer dataset and the more sophisticated feature engineering and ensemble design of the present work. Notably, the present study did not require SMOTE to achieve superior minority-class performance, contrary to the approach of Patel et al.

Fatima et al. evaluated five classifiers (Random Forest, Decision Tree, SVM, KNN, and Logistic Regression) on the 7,999-sample, 20-feature version of the same dataset, with Random Forest achieving 92.3% accuracy, 96% precision, 95% recall, and 96% F1-Score. These figures appear to represent weighted-average metrics across both classes rather than minority-class performance, which would explain the seemingly high values given the 7.77:1 imbalance. The present study's weighted-average F1-Score of 0.98 exceeds these figures, and the more stringent minority-class F1-Score of 0.90 provides a more honest account of performance on the underrepresented safe-water class. Furthermore, the present study employs a more comprehensive experimental design including Optuna hyperparameter optimization, ten-fold stratified cross-validation, and threshold calibration—methodological rigour that is absent from the Fatima et al. study.

4.5.3 Dataset 2: Comparison Against Chen et al. (2024) and Elmotawakkil et al. (2025)

Chen et al. (2024) proposed BS-FAMLP, a hybrid model combining bagging, stacking, GBDT, MLP, and a feature-weighted attention mechanism, trained on a large-scale EEA/World Bank groundwater dataset of 188,623 samples with 15 features. The model achieved accuracy of 0.9616, F1 Macro of 0.9589, and AUC of 0.9834. These figures are substantially higher than the present study's Dataset 2 results (F1 Macro 0.8658, AUC-ROC 0.9991). However, this comparison must be contextualized carefully: the Chen et al. dataset is 188 times larger than the Al-Madhoun dataset used here, and a larger training set dramatically reduces the severity of the minority-class learning problem by providing far more genuine minority examples. Operating on only 1,000 samples with a 152.5:1 imbalance is a fundamentally harder problem. The present study's AUC-ROC of 0.9991 actually surpasses Chen et al.'s 0.9834, suggesting superior probabilistic discrimination even if the absolute F1 Macro is lower—a direct consequence of the extreme class ratio.

Elmotawakkil et al. (2025) proposed HydroPredictor, a CatBoost–MLP hybrid with Optuna tuning and a meta-classifier fusion layer, evaluated on a 315-sample, 3-class Moroccan basin dataset—a sample size even smaller than the Al-Madhoun dataset. Their model achieved accuracy of 0.8923, F1 Macro of 0.8937, Balanced Accuracy of 0.8923, MCC of 0.8394, and AUC > 0.90. The present study's Phase 2 CatBoost on Dataset 2 achieves a higher accuracy (0.9900), higher MCC (0.9352), and higher AUC-ROC (0.9991), while the F1 Macro of 0.8658 is comparable to Elmotawakkil et al.'s 0.8937. Given that the

Al-Madhoun dataset has a substantially more severe class imbalance (152:1 vs. the Feija Basin dataset's estimated class distribution), the present results represent a competitive and, in several metrics, superior performance.

4.5.4 Dataset 3: Comparison Against Wang et al. (2025)

Wang et al. (2025) proposed FreFilterTST, a frequency-domain Transformer architecture employing FFT-based spectral patching, Mixture-of-Experts gating, and inverse FFT reconstruction, specifically designed for unsupervised anomaly detection in the GECCO 2019 drinking water dataset. Their model achieved an Affiliation F1-Score of 0.912 and AUC-ROC of 0.966. The present study's Phase 1 Extra Trees achieves an F1-Score of 0.9514 and AUC-ROC of 0.9976, and the Soft-Voting Ensemble achieves an F1-Score of 0.9295 and AUC-ROC of 0.9986.

This comparison is particularly significant. The FreFilterTST model was specifically architected for temporal anomaly detection, employing sophisticated frequency-domain processing at a parameter count of 2.8 million and a computational cost of 0.12 GFLOPs, and operates in an unsupervised setting (without labelled anomaly examples during training). The present study achieves superior or comparable F1 and AUC metrics using a supervised ensemble of standard tree-based models, trained on temporally engineered tabular features. This demonstrates that domain-informed feature engineering—specifically the extraction of rolling statistics and rate-of-change signals—can bridge the gap between specialized deep learning architectures and general-purpose machine learning methods, achieving state-of-the-art performance with substantially lower model complexity and greater interpretability.

The cross-study comparison yields three overarching conclusions. First, the present study matches or exceeds all comparable state-of-the-art results on Dataset 3 (GECCO), achieving this through principled feature engineering rather than architectural novelty. Second, on Dataset 2, the present study surpasses Elmotawakkil et al. on most absolute metrics despite a substantially more severe class imbalance in the source data, while appropriately falling below Chen et al. who benefit from a dataset two orders of magnitude larger. Third, on Dataset 1, the present study substantially improves upon both Patel et al. and Fatima et al., demonstrating that Optuna-optimised stacking ensembles with engineered features provide a stronger solution than single-model Random Forest baselines with or without SMOTE.

4.6 Conclusion

This chapter presented a thorough evaluation of the proposed ensemble machine learning frameworks on several water quality classification datasets under different imbalance levels and classification settings. The experimental results showed that the fusion of advanced ensemble machine learning algorithms and domain-based feature engineering enabled the development of robust and highly accurate predictive systems for water quality evaluation.

For Dataset 1, excellent binary classification results were achieved, with the Stacking Ensemble emerging as the best-performing model. The experiments on classification performance in the presence of imbalance revealed that SMOTETomek does not bring improvement—actually producing some degradation in performance. This suggests that on moderately imbalanced datasets, complex gradient-boosted ensemble techniques including CatBoost, XGBoost, and LightGBM automatically cope well with imbalance in

the presence of optimally determined thresholds and informative features.

On Dataset 2, which suffered a much more severe imbalance in a multi-class context, good performance was still observed by the proposed framework. CatBoost and ensemble models proved robust and provided high Balanced Accuracy and reasonable AUC-ROC, indicating reasonable separation between different pollution severity levels. The limited sample size and extreme imbalance caused some issues, particularly for the very sparse pollution category. SMOTE-ENN helps to oversample the minority classes and improve recall, but cannot replace the absence of original samples—revealing one of the major challenges in learning with severely imbalanced environmental data.

Several strengths of the approach can be observed. The ensemble approach provides good generalization, stability, and robustness compared to single learners. Boosting algorithms are effective in focusing on difficult samples and exploiting the non-linear structure of physicochemical water quality parameters. The Optuna hyperparameter optimization framework is beneficial in fine-tuning near-optimal configurations with great performance improvement. The feature engineering approach provides an efficient means of integrating domain knowledge and is powerful enough to extract hidden interactions between pollutants not captured directly by individual features.

However, there are also certain limitations. Synthetic oversampling techniques such as SMOTE and combined methods can potentially produce unrealistic data in engineered feature spaces, which may introduce noise and reduce generalizability. Extreme imbalance or absence of data can restrict uniform performance across all categories of the multi-class problem. The high computational cost of ensemble approaches and hyperparameter optimization led to higher training times and greater hardware requirements compared to single-model approaches.

Finally, the conducted study confirms that a fusion of ensemble machine learning techniques and domain-driven feature engineering provides a reliable strategy for intelligent water quality classification. High prediction accuracy, relatively robust performance, and reliable detection of minority classes for imbalanced water quality datasets were observed throughout the proposed approach.

General Conclusion

This thesis explored the application of machine learning and artificial intelligence techniques for water quality prediction, groundwater assessment, and anomaly detection using three different environmental datasets representing tabular classification, groundwater evaluation, and multivariate time-series analysis. The study aimed to investigate how feature engineering, ensemble learning, imbalance handling, and optimization strategies influence predictive performance in environmental monitoring systems.

The experimental results demonstrated that ensemble-based machine learning methods, particularly Random Forest, XGBoost, CatBoost, LightGBM, Extra Trees, and stacking ensembles, provide highly effective solutions for water-related prediction tasks. Across the three datasets, the proposed methodologies achieved strong classification and anomaly-detection performance while maintaining lower computational complexity compared with many deep-learning and hybrid architectures reported in recent literature.

A major finding of this research is that carefully designed feature engineering significantly improves model performance. The introduction of engineered physicochemical interactions, temporal descriptors, statistical indicators, and anomaly-related transformations enabled the models to capture complex environmental relationships more effectively. The results further showed that optimized tree-based ensembles can rival or surpass sophisticated deep-learning systems when supported by informative and domain-driven feature representations.

The thesis also highlighted the critical role of imbalance handling strategies in environmental datasets. While synthetic oversampling methods such as SMOTE are widely used in previous studies, the experiments revealed that oversampling may sometimes degrade performance by generating unrealistic synthetic environmental samples. In several scenarios, native cost-sensitive learning and threshold calibration provided more stable and reliable results than aggressive resampling techniques.

In addition, the systematic use of Bayesian optimization through Optuna improved model generalization and allowed efficient hyperparameter tuning across all experimental phases. The study therefore contributes not only higher predictive performance but also a more rigorous and reproducible machine-learning framework for environmental applications.

Key Contributions and Practical Implications

The principal contributions of this thesis can be summarized as follows:

1. Development of a unified machine-learning framework applied across three heterogeneous environmental datasets.

2. Comprehensive comparative analysis of ensemble learning, hybrid models, and anomaly-detection approaches for water quality monitoring.

3. Systematic investigation of imbalance handling strategies, including SMOTE-family methods and native cost-sensitive learning.

4. Demonstration that feature engineering is often more influential than architectural complexity in environmental prediction tasks.

5. Integration of Bayesian hyperparameter optimization and threshold calibration into all experimental pipelines.

6. Empirical evidence showing that optimized tree ensembles with engineered temporal features can outperform specialized deep-learning models on the **GECCO 2019** benchmark.

7. Detailed methodological evaluation emphasizing interpretability, computational efficiency, and reproducibility.

From a practical perspective, the findings of this thesis may support the development of intelligent environmental monitoring systems capable of assisting water-resource management, pollution detection, public-health protection, and early warning infrastructures. The proposed approaches are computationally efficient and therefore suitable for deployment in real-world monitoring environments with limited hardware resources.

Future Research Directions

Although the proposed framework achieved strong results, several research directions remain open for future investigation:

- Integration of real-time IoT sensor streams for continuous environmental monitoring.

- Development of explainable AI frameworks to improve transparency and trust in environmental decision-making.

- Investigation of uncertainty-aware learning models and probabilistic prediction systems.

- Exploration of graph neural networks and transformer-based architectures for large-scale environmental systems.

- Fusion of multimodal data sources such as satellite imagery, meteorological records, and hydrological simulations.

- Study of online learning and adaptive anomaly-detection systems capable of handling sensor drift and changing environmental conditions.

- Validation of the proposed methods on geographically diverse and industrial-scale datasets.

In closing, this thesis successfully bridges machine learning theory and environmental monitoring applications by demonstrating that robust feature engineering, optimized ensemble learning, and principled imbalance handling can produce highly accurate, interpretable, and computationally efficient water-quality prediction systems. The findings contribute both methodological insights and practical solutions toward the advancement of intelligent environmental monitoring technologies capable of supporting sustainable water-resource management and pollution prevention in real- world scenarios.

Bibliography

[1] Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Arifullah, Alrefaei, A. F., and Albeshr, M. F. (2024). Machine learning models for water quality prediction: A comprehensive analysis and uncertainty assessment in Mirpurkhas, Sindh, Pakistan. *Water*, 16(7):941.

[2] Barnett, V. and Lewis, T. (1994). Outliers in statistical data.

[3] Behmel, S., Damour, M., Ludwig, R., and Rodriguez, M. J. (2016). Water quality monitoring strategies — A review and future perspectives. *Science of The Total Environment*, 571:1312–1329.

[4] Bhateria, R. and Jain, D. (2016). Water quality assessment of lake water: A review. *Sustainable Water Resources Management*, 2(2):161–173.

[5] Boretti, A. and Rosa, L. (2019). Reassessing the projections of the World Water Development Report. *npj Clean Water*, 2(1):15.

[6] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*.

[7] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[8] Campanale, C., Massarelli, C., Savino, I., Locaputo, V., and Uricchio, V. F. (2020). A detailed review study on potential effects of microplastics and additives of concern on human health. *International Journal of Environmental Research and Public Health*, 17(4):1212.

[9] Carvalho, L., Mackay, E. B., Cardoso, A. C., Baattrup-Pedersen, A., Birk, S., Blackstock, K. L., Borics, G., Borja, A., Feld, C. K., Ferreira, M. T., Globevnik, L., Grizzetti, B., Hendry, S., Hering, D., Kelly, M., Langaas, S., Meissner, K., Panagopoulos, Y., Penning, E., Rouillard, J., Sabater, S., Schmedtje, U., Spears, B. M., Venohr, M., Van De Bund, W., and Solheim, A. L. (2019). Protecting and restoring Europe's waters: An analysis of the future development needs of the Water Framework Directive. *Science of The Total Environment*, 658:1228–1238.

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[11] Chen, W., Xu, D., Pan, B., Zhao, Y., and Song, Y. (2024). Machine learning-based water quality classification assessment. *Water*, 16(20):2951.

[12] Cloern, J. E. (2001). Our evolving conceptual model of the coastal eutrophication problem. *Marine Ecology Progress Series*, 210:223–253.

[13] Elmotawakkil, A., Moumane, A., Zahi, A., Sadiki, A., Al Karkouri, J., Batchi, M., Bhagat, S. K., and Enneya, N. (2025). HydroPredictor: A hybrid machine learning model for addressing data scarcity in groundwater prediction. *Scientific Reports*, 15(1):44069.

[14] Fatima, M., Aghzar, M. Z., Ullah, R., and Raza, M. A. (2022). Machine learning-based framework for water quality prediction. *Machine Learning*, 111:123–135.

[15] Fu, G., Jin, Y., Sun, S., Yuan, Z., and Butler, D. (2022). The role of deep learning in urban water management: A critical review. *Water Research*, 223:118973.

[16] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[17] Horton, A. A., Walton, A., Spurgeon, D. J., Lahive, E., and Svendsen, C. (2017). Microplastics in freshwater and terrestrial environments: Evaluating the current understanding to identify the knowledge gaps and future research priorities. *Science of The Total Environment*, 586:127–141.

[18] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[19] Kaddoura, S. (2022). Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability*, 14(18):11478.

[20] Kim, J. et al. (2020). Machine learning approaches for water quality prediction and classification. *Science of the Total Environment*.

[21] Martínez-Villaseñor, L., Vázquez, R. A., and Ochoa-Ruiz, G., editors (2025). *Advances in Soft Computing. MICAI 2025 Posters Track: 24th Mexican International Conference on Artificial Intelligence, MICAI 2025, Guanajuato, Mexico, November 3–7, 2025, Proceedings*, volume 2712 of *Communications in Computer and Information Science*. Springer Nature Switzerland, Cham.

[22] Montgomery, D. C. (2012). Introduction to statistical quality control.

[23] Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S., and Ratna, R. (2022). A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. *Computational Intelligence and Neuroscience*, 2022:1–15.

[24] Qin, B., Yang, L., Chen, F., Zhu, G., Zhang, L., and Chen, Y. (2006). Mechanism and control of lake eutrophication. *Chinese Science Bulletin*, 51(19):2401–2412.

[25] Rezk, N. G., Alshathri, S., Sayed, A., and El-Din Hemdan, E. (2024). EWAIS: An ensemble learning and explainable AI approach for water quality classification toward IoT-enabled systems. *Processes*, 12(12):2771.

[26] Shi, R. (2024). Research on evaluating water pollution determinants using multiple logistic regression. *SHS Web of Conferences*, 208:02016.

[27] Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications*. Springer.

[28] Song, K. (2011). Water quality monitoring using Landsat Thematic Mapper data with empirical algorithms in Chagan Lake, China. *Journal of Applied Remote Sensing*, 5(1):053506.

[29] Taylor, R. G., Scanlon, B., Döll, P., Rodell, M., Van Beek, R., Wada, Y., Longuevergne, L., Leblanc, M., Famiglietti, J. S., Edmunds, M., Konikow, L., Green, T. R., Chen, J., Taniguchi, M., Bierkens, M. F. P., MacDonald, A., Fan, Y., Maxwell, R. M., Yechieli, Y., Gurdak, J. J., Allen, D. M., Shamsudduha, M., Hiscock, K., Yeh, P. J.-F., Holman, I., and Treidel, H. (2013). Ground water and climate change. *Nature Climate Change*, 3(4):322–329.

[30] Tsay, R. S. (2005). Analysis of financial time series.

[31] Tukey, J. W. (1977). Exploratory data analysis.

[32] Wang, S. et al. (2019). Water quality assessment and prediction using machine learning: a review. *Environmental Modelling Software*.

[33] Wang, Y., Zhang, J. J., and Zhang, M. Y. (2025). FreFilterTST: A dynamic channel graph sparsification approach to multivariate time series anomaly detection with frequency-domain restoration. *Scientific Reports*, 16(1):3289.

[34] World Health Organization (2022). *Guidelines for Drinking-Water Quality*. World Health Organization, Geneva, 4th, incorporating the first and second addenda edition.

[35] World Health Organization and United Nations Children's Fund (2023). Progress on household drinking water, sanitation and hygiene 2000–2022: Special focus on gender. Technical report, WHO & UNICEF, Geneva.

[36] Xia, Y., Zhang, M., Tsang, D. C. W., Geng, N., Lu, D., Zhu, L., Igalavithana, A. D., Dissanayake, P. D., Rinklebe, J., Yang, X., and Ok, Y. S. (2020). Recent advances in control technologies for non-point source pollution with nitrogen and phosphorous from agricultural runoff: Current practices and future prospects. *Applied Biological Chemistry*, 63(1):8.

[37] Xin, L. and Mou, T. (2022). Research on the application of multimodal-based machine learning algorithms to water quality classification. *Wireless Communications and Mobile Computing*, 2022:9555790.

[38] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., and Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, 1(2):107–116.