Bhuminkumar Parmar (X0887875)

MACHINE LEARNING IN WATER TREATMENT PLANTS FOR WATER QUALITY PREDICTION AND PREDICTIVE MAINTENANCE

Master's Programme in Computing Science,
Sustainable and Autonomous Systems
Supervisor: Prof. Mohammed Elmusrati
Instructor: Prof. Petri Välisuo

Vaasa 2026

UNIVERSITY OF VAASA
School of science and technology
Author: Bhuminkumar Parmar (X0887875)
Title of the thesis: Machine Learning in Water Treatment Plants for Water Quality Prediction and Predictive Maintenance
Degree: Master of Science in Technology
Degree Programme: Master's Programme in Computing Science
Major: Sustainable and Autonomous Systems
Supervisor: Prof. Mohammed Elmusrati
Instructor Prof. Petri Välisuo
Year: 2026 **Pages:** 118

ABSTRACT

Water treatment plants are critical infrastructure for public health and sustainable resource management. Recently, Data center is a major industry which consume large source of ultra-pure water for cooling. Most conventional water treatment facilities continue to operate using rule-based control systems and reactive monitoring practices, which limit their capacity for early detection of water quality deterioration and proactive maintenance planning. In Finland, growing national emphasis on digital transformation and sustainable development creates compelling opportunity to enhance an existing water treatment infrastructure with introducing machine learning through digital retrofitting rather than costly replacement.

This thesis develops and evaluates a machine learning-based framework for water quality prediction by water chemical parameters and base on that predictive maintenance in conventional water treatment plants. The research develops ML models to predict water potability, compare algorithms, classify samples, create KNN health index, and enable retrofits without major infrastructure changes.

This study utilizes the Water Potability data-set from Kaggle , comprising over three thousand water samples which is characterized by nine physico-chemical parameters including pH, sulfate, hardness, turbidity, chloramines, and trihalomethanes. The data preprocessing pipeline incorporates class-wise median imputation for missing values, inter-quartile range-based outlier capping, StandardScaler feature scaling, and SMOTE oversampling to address class imbalance. After that, seven supervised classification algorithms are implemented and tuned using GridSearchCV with five-fold cross-validation which are: Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors, and AdaBoost.

The results demonstrate that ensemble tree-based methods consistently outperform distance-based and linear classifiers for water potability classification. Random Forest achieves the highest performance score , with an accuracy of approximately 79 % and an area under the ROC curve of approximately 0.88. The algorithm performance ranking is consistent with published comparative literature, strengthening confidence in its venerability. Feature importance analysis identifies sulfate and pH as the two dominant predictive features, together accounting for nearly half of the model's predictive capacity. Turbidity, despite its operational prominence, ranks the lowest in this dataset due to its limited variation across all samples.

The KNN-based Health Index framework assigns a continuous health score from zero to one hundred to each water sample based on its proximity to a WHO guideline-defined ideal reference state.

The framework classifies water samples into three operational maintenance zones: Healthy (HI ≥ 70), Warning (40–69), and Critical (HI < 40). A 365-day RO membrane maintenance simulation demonstrates that Health Index-triggered CIP scheduling reduces cleaning cycles by approximately 20 percent compared to conventional fixed-schedule maintenance - 20 versus 25 CIP events per year while keeping the membrane continuously within the Healthy and Warning zones. Traditional fixed-schedule maintenance, by contrast, performs approximately 30 percent of CIP events unnecessarily in the Healthy zone, wasting chemical resources, and 10 percent re-actively in the Critical zone, where increased energy consumption and membrane damage have already occurred. This finding is consistent with Gradiant's SmartOps AI deployment at the Bedok NEWater Factory, Singapore, where condition-based ML maintenance achieved 98.1 percent cleaning prediction accuracy (Gradiant, 2024).

The integrated framework operates exclusively on parameters available from existing monitoring infrastructure, confirming its technical feasibility as a digital retrofitting for conventional water treatment plants. The findings have specific relevance for Finnish water utilities operating under the Health Protection Act and the EU Drinking Water Directive. The thesis contributes a validated, open-source, reproducible methodological framework that advances the state of the art in data-driven water treatment management and provides a practical foundation for the adoption of machine learning-based decision support in water utility operations.

Keywords: machine learning, water quality prediction, water treatment plants, predictive maintenance, health index, Random Forest, digital retrofitting, potability classification

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
AUC	Area Under the Receiver Operating Characteristic Curve
CIP	Clean-in-place
DAF	Dissolved Air Flotation
DBP	Disinfection By-Product
DOC	Dissolved Organic Carbon
EC	Electrical Conductivity
EU	European Union
F1	F1-Score (Harmonic Mean of Precision and Recall)
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory Network
LR	Logistic Regression
MF	Microfiltration
ML	Machine Learning
NF	Nanofiltration
NOM	Natural Organic Matter
NTU	Nephelometric Turbidity Units
PID	Proportional-Integral-Derivative (Controller)
PLC	Programmable Logic Controller
RF	Random Forest
RO	Reverse Osmosis
ROC	Receiver Operating Characteristic
RUL	Remaining Useful Life
SCADA	Supervisory Control and Data Acquisition
SDG	Sustainable Development Goal
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
SYKE	Finnish Environment Institute (Suomen ympäristökeskus)
TDS	Total Dissolved Solids
THM	Trihalomethane
TOC	Total Organic Carbon
UF	Ultrafiltration
UV	Ultraviolet
WHO	World Health Organization
WSP	Water Safety Plan

WTP	Water Treatment Plant
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

Back to Contents

1 INTRODUCTION

1.1 Background and Motivation

Water is the most fundamental resource for human survival, and ensuring its safety and quality is one of the central responsibilities of modern public infrastructure. Water treatment plants (WTPs) serve as the critical interface between raw water sources and safe drinking water delivered to communities. In Finland, as in many parts of the world, these facilities process millions of cubic meters of water annually , relying on complex sequences of physical, chemical, and biological treatment stages to meet stringent water quality standards.

Despite their critical importance, most conventional water treatment plants continue to operate using rule-based control systems and reactive monitoring practices. Operators typically depend on periodic laboratory analyses, manual inspections, and experience-based decision-making to assess water quality and schedule maintenance activities. While these approaches have proven effective for decades, they have inherent limitations in an era of optimizing energy use and chemical waste, increasing operational complexity, aging infrastructure, and tightening resource constraints.

In Nordic countries, Elevated chloride concentrations during winter season may impair the reliability of waste water analysis. The chloride salt used in anti-slip treatment is transported into sewer network with runoff water from roads and streets area. During winter season and specially during periods when temperature fluctuates on both side of zero degree, the chloride; turbidity and other chemical parameters usually rise of a sudden or short period of time. If this changed water conditions are not taken into account, measurement results may be distorted. This parameters are measured by laboratory analysis which takes time. So In that time input water quality can deteriorate UF and RO membranes performance even for short period of time. That’s where machine learning helps. By past data analysis, preventive maintenance of plant can be confirmed.

Back to Contents

In Finland, there is a growing national emphasis on digital transformation and sustainable development of all public utilities. The Finnish government and the environmental agencies have increasingly called for smarter, data-driven solutions to manage water resources more efficiently. This context presents a compelling opportunity to explore the application of machine learning (ML) in water treatment operations — not through costly infrastructure replacement, but through intelligent AI based digital retrofitting of existing systems.

Machine learning offers a fundamentally different approach to water quality management. Unlike the static rule-based/expert systems, ML models learn complex, nonlinear relationships directly from historical operational data. They are able to identify subtle patterns that precede water quality violations or equipment failures -- patterns that would not be foreseeable in conventional monitoring approaches. Through integrating ML into existing water treatment infrastructure, utilities can transition from reactive to more predictive operations, giving early warnings, optimized chemical dosing , and proactive maintenance scheduling.

This thesis studies the application of machine learning techniques in water treatment plants, focusing on two interrelated problems : water quality prediction and classification, and predictive maintenance using a health index approach. The research is positioned within the broader context of sustainable and autonomous systems engineering systems, in accordance with the University of Vaasa's focus on sustainable industrial transformation.

1.2 Research Problem :

Despite significant advances in industrial fully automated systems and data analytics, the water treatment sector has been relatively slow to adopt machine learning technologies. Many water utilities - particularly medium and smaller-scale water treatment plants — still rely on manual sampling, delayed laboratory results, and scheduled fixed days maintenance intervals that are not responsive to actual system conditions. This reactive paradigm creates several critical operational challenges.

Back to Contents

First, the delayed detection of water quality deterioration poses a direct risk to public health if it is for human usage. By the time a quality violation is identified through conventional sampling, contaminated water may have already progressed through the distribution network. Second, the absence of predictive maintenance frameworks leads to unplanned equipment failures - particularly in filtration systems and reverse osmosis (RO) membranes — resulting in costly downtime and emergency repairs. Third, the underutilization of historical operational data represents a significant missed opportunity; as water treatment plants continuously generate large volumes of sensor data, yet this data is rarely used for predictive modelling.

Existing research has studied machine learning applications in water quality prediction. But still there are gaps remain in several important areas. Most studies focus either on water quality prediction or maintenance planning in isolation, rather than presenting an integrated framework. Furthermore, few studies address the specific challenge of retrofitting conventional plants - which is deploying ML solutions within existing infrastructure without demanding significant hardware upgrades or sensor installations.

This research addresses these all gaps by developing and evaluating a comprehensive ML-based framework that encompasses both water quality classification and predictive maintenance , designed specifically for use in conventional water treatment plants operating in a Finnish and Nordic countries context.

1.3 Research Objectives and Questions

The primary aim of this thesis is to develop and evaluate machine learning-based models for water quality prediction and predictive maintenance, enabling the retrofitting of conventional water treatment plants towards sustainable and semi-autonomous operation.

Back to Contents

The specific research objectives are:

(1) To develop and test machine learning models for predicting water quality using key chemical parameters including pH, turbidity, conductivity, hardness and others.

(2) To compare the performance of multiple machine learning algorithms - including Support Vector Machine, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors, Logistic Regression, and AdaBoost in water quality classification tasks.

(3) To classify water samples as potable or non-potable based on measured water quality parameters, and evaluate classification reliability.

(4) To develop a machine learning-based Health Index using the K-Nearest Neighbors algorithm for predictive maintenance of filtration and RO membrane systems.

(5) To investigate the feasibility of deploying ML-based retrofitting solutions in conventional water treatment plants without major infrastructure modifications.

These objectives are guided by the following research questions:

RQ1: How accurately can machine learning models predict water quality using commonly measured chemical parameters?

RQ2: Which machine learning algorithms perform best in water quality prediction and classification tasks, and what factors explain performance differences?

RQ3: Can machine learning-based classification reliably distinguish between potable and non-potable water to a standard suitable for operational decision support?

RQ4: How is it possible for a machine learning-based health index support predictive maintenance decisions in water treatment plants?

RQ5: How successfully can machine learning be used to retrofit conventional water treatment plants without major infrastructure facility changes?

Back to Contents

1.4 Scope and Limitations

This research focuses on the application of supervised machine learning techniques to water quality data derived from publicly available datasets. The study scope is defined by the following boundaries and assumptions.

In terms of data, the research utilizes the Water Potability dataset obtained from Kaggle , which contains 3,276 water samples characterised by nine physicochemical parameters: pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The dataset includes a binary potability label, enabling classification model development and evaluation.

The scope of machine learning methods is limited to supervised classification algorithms. Deep learning architectures and unsupervised methods are outside the scope of this study. Also, The predictive maintenance component is developed as a proxy framework — using water quality parameters as indirect indicators of system health — rather than being based on direct equipment condition monitoring data, which was not available for this study.

The research does not include a real-time deployment or integration with actual plant control systems. The proposed framework is evaluated in a simulation environment and to demonstrate the approach's conceptual and technical feasibility. The Generalization of results to all water treatment plants and water source types requires further validation with proper site-specific operational data.

These limitations do not diminish the scientific contribution of the work; rather, they define the boundaries within which the findings should be interpreted and applied. The framework developed provides a validated foundation for more operationally integrated solutions.

1.5 Structure of the Thesis

Back to Contents

This thesis is organized into seven chapters, each addressing a distinct component of the research.

Chapter 1 (Introduction) presents the background and motivation for the research, defines the research problem, states the research objectives and questions , delineates the scope and limitations, and outlines the structure of the thesis.

Chapter 2 (Water Treatment Plants: Processes and Challenges) provides technical background of conventional water treatment processes, quality assessment methods, and the operational challenges that motivate the adoption of machine learning solutions.

Chapter 3 (Literature Review) reviews the existing research on machine learning applications in water quality monitoring and prediction, classification approaches for potability assessment, and predictive maintenance frameworks in industrial systems. The chapter identifies research gaps that justify the present study.

Chapter 4 (Research Methodology) describes the data collection and preprocessing procedures, the machine learning models implemented, and the health index development methodology, and the evaluation metrics and validation strategies used in the study.

Chapter 5 (Results and Analysis) presents the empirical findings of the study, including water quality classification performance across all seven models, feature importance analysis, and health index results with maintenance zone classification.

Chapter 6 (Discussion) interprets the results in light of the research questions, discusses a theoretical and practical implications, and situates the findings within the broader literature on digital transformation of water utilities.

Chapter 7 (Conclusions) summarize the key contributions of the thesis, provides recommendations for water utility practitioners and policymakers, and outlines directions for future research.

Back to Contents

2 WATER TREATMENT PLANTS: PROCESSES AND CHALLENGES

This chapter provides the technical and operational background of plants which is necessary to contextualize the machine learning framework developed in this thesis. Section 2.1 introduces the role and significance of water treatment plants in public health and environmental management. Section 2.2 describes the principal stages of conventional water treatment processes. Section 2.3 reviews the key water quality parameters used in this study and their significance for operations work. Section 2.4 examines current monitoring and control practices and Section 2.5 identifies the operational challenges that is motivating the adoption of data-driven solutions.

2.1 Role and Significance of Water Treatment Plants

Water treatment plants are engineered facilities designed to remove organic and inorganic matter and contaminants from raw water sources - including surface water from rivers and lakes, and groundwater from aquifers -- to produce water that meets regulatory quality standards for human consumption and environmental discharge. Their operation is fundamental to public health, since unsafe drinking water remains one of the leading causes of preventable illness globally (World Health Organization [WHO], 2022).

In Finland, drinking water quality is regulated under Health Protection Act (763/1994) and its associated decrees, which are aligned with the European Union Drinking Water Directive (2020/2184/EU). Finnish water utilities are required to deliver water that meets strict physicochemical and microbiological standards, including pH, turbidity, nitrate levels, and coliform bacteria counts, among others. The Finnish Environment Institute (SYKE) reports that approximately 90% of Finnish municipal drinking water is sourced from surface waters, predominantly from lakes, while the remaining supply is drawn from groundwater sources (SYKE, 2021).

Water treatment plants in Finland and other Nordic countries face distinctive operational conditions. Raw water sources are characterized by low mineral content

Back to Contents

and low concentrations of natural organic matter, but are susceptible to seasonal variations in temperature, post- rain fall turbidity, and algal blooms during warmer months. These fluctuations present ongoing challenges for the treatment process stability and water quality consistency.

Beyond their public health function, water treatment plants carry significant economic and environmental responsibilities. Energy consumption in water treatment and distribution represents a major operational cost, and the treatment chemicals used — including coagulants, disinfectants, and pH adjusters have associated environmental impacts that utilities always seek to minimize. The dual imperative of maintaining the water quality while reducing resource consumption makes intelligent, data-driven process management an increasingly attractive proposition for water utility operators.

2.2 Conventional Water Treatment Processes

The treatment of raw water to potable standard involves a series of interdependent unit processes, each targeting specific categories of contaminants. While the precise configuration of a treatment plant depends on the characteristics of the source water and the required product quality,most conventional surface water treatment plants employ the following principal stages.

2.2.1 Coagulation and Flocculation :

Coagulation and Flocculation constitute the first chemical treatment stage in most conventional water treatment plants. Coagulation involves the addition of chemical agents - typically aluminium sulfate (alum or iron-based coagulants (FECL3) to destabilize the colloidal particles in the raw water. These colloidal particles, which include clay, silt, organic matter, and microorganisms, carry negative surface charges that keep them dispersed in the water column. Coagulants neutralize these charges, allowing particles to aggregate.

Flocculation follows coagulation and involves gentle agitation of the water to promote the growth of larger particle clusters, known as flocs. The efficiency of coagulation and

Back to Contents

flocculation is highly sensitive to raw water characteristics, including pH, temperature, turbidity, and the natural organic matter concentration. Optimal chemical dosing at this stage propagates inefficiencies through all subsequent treatment stages, underscoring the importance of real-time process monitoring and adaptive control.

2.2.2 Sedimentation

Following flocculation, the water enters sedimentation basins where the formed flocs settle under gravity, separating from the treated water. Sedimentation efficiency depends on the size and density of the floc particles, water temperature, and hydraulic loading rates. In modern plants, lamella settlers or dissolved air flotation (DAF) systems may supplement or replace conventional sedimentation, particularly for low-turbidity source waters or water with high concentrations of low-density organic particles.

The sludge accumulated at the base of sedimentation tanks requires periodic removal and disposal, representing a significant operational and environmental management consideration. Sludge handling costs can constitute a notable fraction of total plant operating expenses and their volume and composition are directly influenced by the upstream coagulation process.

2.2.3 Filtration

Filtration removes residual suspended particles and microorganisms that pass through the sedimentation stage. Conventional granular media filters (MGF,DMF,ACF) typically comprising layers of sand, anthracite, or gravel that trap particles through a combination of mechanical straining, adsorption, and biological activity within the filter bed. Rapid sand filters are the most commonly employed configuration in large-scale municipal water treatment, which overall reduces the turbidity of water.

Membrane filtration technologies, including microfiltration (MF), ultrafiltration (UF), nanofiltration (NF), and reverse osmosis (RO) are increasingly adopted as alternatives or supplements to granular media filtration. Membrane systems offer superior

Back to Contents

removal of pathogens, dissolved organic compounds, and micro-pollutants, but require careful maintenance to manage scaling and fouling - the accumulation of deposits on membrane surfaces that progressively reduces permeate flux and increases operating pressure.

Membrane scaling and fouling are one of the primary operational challenges in advanced water treatment plants and its management is a direct motivation for the predictive maintenance framework developed in this thesis. Early detection of fouling onset, based on indirect water quality indicators, can enable timely cleaning interventions that extend membrane service life and reduce energy consumption.

2.2.4 Disinfection

Disinfection is the final and most critical barrier against waterborne pathogens in the treatment process. Chlorination remains the most widely used disinfection method globally, owing to its effectiveness, low cost, and residual protection within the distribution network. Chloramines , formed by the reaction of chlorine with ammonia are increasingly used as a secondary disinfectant, particularly in systems where the formation of trihalomethane (THM) disinfection by-products is a concern.

Trihalomethanes and haloacetic acids are formed when chlorine-based disinfectants react with natural organic matter present in the water. These disinfection by-products are subject to regulatory limits in Finland and the European Union due to their potential carcinogenicity at elevated concentrations. The management of THM formation therefore requires a careful balancing of disinfection against by-product control a trade-off that is dependent on real-time knowledge of organic carbon levels and other water quality parameters.

Alternative disinfection technologies, including ultraviolet (UV) irradiation and organization, are employed at many modern facilities, often in combination with chlorination to provide both primary inactivation and residual protection. The selection and optimization of disinfection strategies represent an important

Back to Contents

operational decisions that can benefit from data-driven monitoring and predictive modelling .

2.2.5 pH Adjustment and Post-Treatment

Following disinfection, the pH of treated water is typically adjusted to reduce corrosivity and protect distribution infrastructure. Lime or sodium hydroxide is commonly added to raise pH to the range of 7.5 - 9.5, which minimizes the leaching of metals such as lead and copper from pipe materials. Phosphate compounds may also be added to form a protective coating on pipe surfaces, further reducing corrosion.

In some treatment configurations, activated carbon filtration or advanced oxidation processes are employed as additional post-treatment stages to remove trace organic micro pollutants, taste and odour compounds or pharmaceutical residues. The increasing presence of emerging contaminants in water sources has driven growing interest in these advanced treatment technologies.

Table 1.Summary of conventional water treatment stages and their primary functions.

Treatment Stage	Primary Mechanism	Target Contaminants	Key Parameters
Coagulation & Flocculation	Chemical destabilisation	Colloids, suspended solids, NOM	pH, turbidity, coagulant dose
Sedimentation	Gravity settling	Floc particles, suspended solids	Hydraulic retention time, turbidity
Filtration (Granular)	Mechanical straining, adsorption	Residual particles, microorganisms	Head loss, turbidity, flow rate
Membrane Filtration (RO/UF)	Size exclusion, pressure-driven	Dissolved solids, pathogens, micropollutants	Transmembrane pressure, flux, fouling
Disinfection	Chemical/UV inactivation	Pathogens, bacteria, viruses	Chloramine dose, pH, contact time, THMs
pH Adjustment	Chemical addition	Corrosivity, metal leaching	pH, alkalinity, calcium hardness

Back to Contents

2.3 Key Water Quality Parameters

Water quality assessment relies on measuring a range of chemical and biological parameters, each providing information on different aspects of water composition and treatment performance. This section describes the nine parameters used in the present study, which constitute a representative subset of those routinely monitored in operational water treatment plants.

2.3.1 pH

pH is the measure of hydrogen ion concentration in water, which is expressed on a logarithmic scale from 0 to 14. It is one of the most fundamental water quality parameters, influencing the solubility of metals, the efficacy of disinfection and the corrosivity of water towards pipe materials. The WHO guideline value for drinking water pH is 6.5 - 8.5, though values between 7.0 and 8.0 are generally preferred for distribution system compatibility (WHO, 2022). Extreme pH values can indicate contamination events, chemical dosing errors or algal activity, making pH monitoring critical for both treatment control and potability assessment.

2.3.2 Hardness

Water hardness is primarily determined by the concentrations of calcium and magnesium ions, expressed as equivalent calcium carbonate ($CaCO_3$) in mg/L. Hard water can cause scaling in pipes and heat exchangers , which is reducing system efficiency and increase maintenance requirements. Very soft water, by contrast, tends to be more corrosive. In the context of water treatment, hardness affects coagulation efficiency and chemical dosing requirements. Finnish surface waters are characteristically soft, typically in the range of 20 - 80 mg/L $CaCO_3$, which contrasts with the global dataset used in this study where hardness values extend to over 300 mg/L.

2.3.3 Total Dissolved Solids

Back to Contents

Total dissolved solids (TDS) represent the aggregate concentration of all dissolved inorganic and organic substances in water, measured in mg/L. TDS is an indicator of the overall mineral content of water and serves as a proxy for water conductivity and overall ionic load. The WHO guideline value for TDS in drinking water is 1,000 mg/L, though values below 500 mg/L are generally preferred for palatability, and values below 300 mg/L are considered very good. Elevated TDS concentrations can indicate contamination from agricultural runoff, industrial discharges, or natural geological sources.

2.3.4 Chloramines

Chloramines are secondary disinfectants formed by the controlled reaction of chlorine with ammonia in treated water. They are measured in mg/L and provide a more stable and longer-lasting residual disinfection effect than free chlorine alone, which makes them particularly valuable in extended distribution systems. However, chloramine concentrations must be carefully controlled. concentrations that are too low leave the distribution system vulnerable to bacterial regrowth, while excessively high concentrations can cause taste and odour problems and contribute to the formation of nitrogenous disinfection by-products.

2.3.5 Sulfate

Sulfate (SO_4^{2-}) occurs naturally in water from the dissolution of sulfate-bearing minerals and from atmospheric deposition. Elevated sulfate concentrations can impart a bitter taste to drinking water and, at very high concentrations (above 500 mg/L), may have a laxative effect on consumers unfamiliar with high-sulfate water. The WHO guideline value for sulfate is 500 mg/L, with an aesthetic objective of 250 mg/L. In the context of the present study, sulfate emerged as the most important predictive feature in the machine learning models, which may reflect its role as an integrative indicator of overall mineral content and source water characteristics.

2.3.6 Conductivity

Back to Contents

Electrical conductivity (EC) measures the ability of water to conduct an electrical current, which is directly related to the concentration of dissolved ionic species. Conductivity is expressed in microsiemens per centimetre (µS/cm) and provides a rapid, non-destructive indication of overall dissolved mineral content. It is commonly used as a surrogate for TDS in continuous online monitoring systems. Sudden changes in conductivity can indicate intrusion events, chemical dosing anomalies or changes in source water quality, making it a valuable parameter for real-time process monitoring.

2.3.7 Organic Carbon

Total organic carbon (TOC) and dissolved organic carbon (DOC) quantify the concentration of organic matter in water, expressed in mg/L. Natural organic matter (NOM) in source water originates from the decomposition of plant and animal material and varies seasonally and with catchment land use. TOC / DOC is a critical operational parameter because organic matter is a precursor to disinfection by-products, particularly trihalomethanes and haloacetic acids formed during chlorination. Elevated organic carbon concentrations also increase coagulant demand and can promote biological regrowth in distribution systems.

2.3.8 Trihalomethanes

Trihalomethanes (THMs) are a group of chemical compounds — including chloroform, bromodichloromethane, dibromochloromethane, and bromoform - formed as by-products of chlorination in the presence of natural organic matter. They are measured in micrograms per litre (µg/L). The WHO guideline value for total THMs in drinking water is 300 µg/L, while the EU Drinking Water Directive (2020/2184/EU) sets a stricter limit of 100 µg/L. THM concentrations in treated water reflect the combined influence of organic carbon content, chlorine dose, contact time, pH, and temperature , making them a complex but informative indicator of disinfection by-product formation potential.

Back to Contents

2.3.9 Turbidity

Turbidity is an optical property of water that measures the scattering of light by suspended particles, expressed in nephelometric turbidity units (NTU). It is one of the most commonly monitored parameters in water treatment, serving as an indicator of particulate contamination, filter performance and potential pathogen presence. The WHO guideline value for turbidity in drinking water is 1 NTU, with values below 0.1 NTU indicating optimal filtration performance. Turbidity spikes in treated water may indicate filter breakthrough events, inadequate coagulation, or sedimentation inefficiencies, all of which have direct implications for treatment system health.

Table 2. Water quality parameters used in this study, with WHO guideline values and significance for treatment operations.

Parameter	Unit	WHO Guideline	Treatment Significance
pH	—	6.5 – 8.5	Coagulation efficiency, corrosivity, disinfection efficacy
Hardness	mg/L CaCO3	No guideline (aesthetic)	Scaling potential, coagulant demand
Total Dissolved Solids	mg/L	< 1,000 (aesthetic)	Overall mineral content, palatability
Chloramines	mg/L	< 3 (as Cl2)	Residual disinfection, DBP formation
Sulfate	mg/L	< 500 (< 250 aesthetic)	Taste, mineral indicator, scaling
Conductivity	µS/cm	No specific limit	Proxy for TDS, intrusion detection
Organic Carbon	mg/L	No specific limit	DBP precursor, NOM indicator
Trihalomethanes	µg/L	< 300 (EU: < 100)	Disinfection by-product, health risk
Turbidity	NTU	< 1 (ideally < 0.1)	Filtration performance, pathogen indicator

Back to Contents

2.4 Current Monitoring and Control Practices

The monitoring and control of water treatment processes has evolved considerably over the past several decades, from entirely manual operations to increasingly automated systems incorporating online sensors, supervisory control and data acquisition (SCADA) infrastructure, and programmable logic controllers (PLCs). However, the degree of automation and sophistication of monitoring practices vary considerably across the global water utility sector.

2.4.1 Online and Laboratory-Based Monitoring :

Modern water treatment plants consists a combination of online (continuous) and offline (laboratory-based) monitoring methods. Online sensors provide real-time measurements of parameters such as pH, turbidity, conductivity, chlorine residual and dissolved oxygen at strategic points throughout the treatment process. These sensors generate high-frequency data streams that are typically logged by SCADA systems, providing operators with a continuously updated picture of process performance.

Offline laboratory analysis is conducted at regular intervals by chemist typically daily or weekly for parameters that cannot yet be measured reliably by online sensors, including detailed microbiological analysis, THM concentrations, heavy metals and pesticide residues. Laboratory results are essential for regulatory compliance reporting but introduce a time delay between sample collection and result availability, which limits their utility for real-time process control.

2.4.2 Rule-Based Control Systems

The control of treatment processes in most conventional plants is implemented through rule-based systems, in which predefined thresholds and set-points trigger specific control actions. For example, a turbidity sensor downstream of a filter may trigger an alarm if readings exceed a defined threshold, prompting an operator to initiate a backwash cycle. Chemical dosing pumps may be controlled by proportional-

Back to Contents

integral-derivative (PID) controllers that adjust dosing rates in response to deviations from target values.

While rule-based systems have proven reliable under stable operating conditions, they exhibit important limitations when faced with complex , nonlinear process dynamics or novel operating conditions outside their design envelope. Fixed thresholds cannot adapt to a gradual changes in source water quality, seasonal variations or the interactive effects of multiple parameters changing simultaneously. Furthermore, rule-based systems are inherently reactive. They respond to deviations that have already occurred rather than anticipating them in advance.

2.4.3 Maintenance Planning Practices

Maintenance practices in water treatment plants traditionally follow one of two approaches : corrective maintenance, in which equipment is repaired or replaced after failure, or preventive maintenance, in which maintenance activities are scheduled at fixed intervals based on manufacturer recommendations or historical experience. Both approaches have significant limitations.

Corrective maintenance results in unplanned downtime, potentially compromising the water supply continuity and quality. This often incurring higher repair costs than planned interventions. Preventive maintenance avoids emergency failures but may result in unnecessary maintenance of equipment that is still in good condition. This also wastes resources and potentially introducing new failure modes through the maintenance process itself. Predictive maintenance - in which the actual condition of equipment is continuously assessed to schedule interventions only when needed - offers a superior alternative that minimizes both unnecessary maintenance and unplanned failures.

2.5 Operational Challenges in Conventional Water Treatment Plants

Back to Contents

The preceding sections have introduced the processes and monitoring practices of conventional water treatment. This section synthesizes the principal operational challenges that motivates the machine learning approach developed in this thesis.

2.5.1 Delayed Detection of Water Quality Deterioration

The reliance on periodic laboratory sampling for key parameters such as THMs, microbiological indicators, and trace contaminants means that water quality deterioration may not be detected for hours or even days after its onset. During this interval, substandard water may be distributed to consumers or discharged to receiving water bodies. Early warning systems capable of detecting incipient quality deterioration in near-real-time are therefore a high priority for water utilities and seek to improve their operational resilience.

2.5.2 Under-utilization of Historical Data

SCADA systems in modern water treatment plants generate enormous volumes of operational data including sensor readings, alarm histories, chemical dosing records, and maintenance logs in detail log sheet. Despite the potential analytical value of this data, it is very likely exploited beyond immediate operational monitoring. The absence of machine learning or advanced statistical tools means that patterns which is predictive of future quality events or equipment failures may be present but it is subtle in the historical record and waste as undetected. This represents a significant missed opportunity for operational improvement.

2.5.3 Membrane Fouling and Filter Performance Degradation

In plants employing membrane filtration or reverse osmosis, membrane fouling is a persistent and costly operational challenge. Fouling - caused by the accumulation of organic matter, inorganic matter, and biological growth on membrane surfaces - progressively reduces water permeability, increases energy consumption, and ultimately requires chemical cleaning or membrane replacement. The onset of fouling

Back to Contents

is gradual and may not be apparent from routine monitoring until significant performance degradation has occurred. A predictive health index capable of tracking membrane condition trends and providing advance warning of fouling could substantially reduce the maintenance costs and extend membrane service life.

2.5.4 Aging Infrastructure and Resource Constraints

A significant proportion of water treatment infrastructure in Finland and other developed countries was constructed in the mid-to-late twentieth century and is approaching or has exceeded its design life even though it is fully or semi automation. Complete infrastructure replacement is economically prohibitive for many utilities. Digital retrofitting - the addition of data analytics capabilities to existing infrastructure offers a cost-effective pathway to improved operational performance without requiring wholesale capital investment. Machine learning solutions that operate on data from existing sensors are particularly well suited to this context.

2.5.5 Climate Change and Source Water Variability

Climate change is intensifying variability in source water quality, with increasing the frequency of extreme rainfall events leading to turbidity and nutrient loading spikes, warmer temperatures promoting algal growth, and changing seasonal patterns affecting the performance of treatment processes calibrated for historical conditions. Adaptive and data-driven process control systems that can respond dynamically to changing input conditions are therefore increasingly important for maintaining consistent treated water quality in the face of a changing climate.

2.6 Chapter Summary

This chapter has provided a comprehensive overview of water treatment plant processes, quality parameters and operational practices. The discussion has highlighted five principal challenges motivating the application of machine learning in this domain : delayed detection of water quality deterioration, under utilization of

Back to Contents

historical data, membrane fouling and filter degradation, aging infrastructure, and increasing source water variability due to climate change.

The nine water quality parameters are central to this study - pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity which is been described in terms of their chemical significance, WHO guideline values, and relevance to treatment operations. Table 1 and Table 2 provides a consolidated reference summaries of treatment stages and quality parameters respectively.

These operational realities establish a clear and compelling case for ML-based retrofitting of conventional water treatment plants. Chapter 3 reviews the existing scientific literature on machine learning applications in this domain, identifies gaps in current knowledge & situates the present research within a broader academic context.

Back to Contents

3 LITERATURE REVIEW

This chapter reviews the scientific literature relevant to the research objectives of this thesis. The review is organized into six thematic sections. Section 3.1 discusses Industry adoption of AI and ML in water treatment operations. Section 3.2 examines machine learning applications in water quality monitoring and prediction. Section 3.3 reviews the classification approaches for potable water assessment. Section 3.4 reviews the ensemble and tree-based methods, which emerged as the best-performing model category in this study. Section 3.5 covers predictive maintenance frameworks and health index concepts in industrial systems. Section 3.6 discusses the specific challenge of ML-based retrofitting of conventional infrastructure. Section 3.7 examines the identified research gaps that justify and position the present study.

3.1 Industry Adoption of AI and ML in Water Treatment Operations

While the academic literature establishes the theoretical and experimental foundations for ML in water treatment, the commercial sector provides the complementary evidence. AI-based operational intelligence has already transitioned from research to large scale industrial deployment. Several leading global water technology companies have developed and commercially deployed ML platforms specifically targeting the predictive maintenance and operational optimization challenges which is addressed in this thesis.

3.1.1 Gradiant SmartOps AI - Condition-Based Membrane Maintenance

Gradiant, a global water technology company operating across more than 600 projects worldwide, has developed the SmartOps AI platform — described by the company as a digital ecosystem for the control, prediction, and performance optimization of water treatment facilities (Gradiant, 2024). The platform creates a digital twin of each facility, with SCADA sensors streaming real-time operational data between the physical plant and its virtual model. Machine learning algorithms analyse this data stream to proactively predict the likelihood of future maintenance events, including the need for RO membrane cleaning and replacement.

Back to Contents

A particularly relevant deployment of SmartOps AI was conducted at the Bedok NEWater Factory in Singapore. It was in collaboration with PUB Singapore's National Water Agency. The objective was to determine the optimal schedule for RO membrane CIP cleaning and replacing the conventional time-based fixed schedule with a data-driven condition-based approach. The ML algorithm analyzed daily normalized differential pressure data to detect the onset of biofouling and categorizing membrane condition into three risk levels - Yellow, Orange, and Red. This levels corresponds to progressive stages of RO fouling. Rigorous testing gave a 98.1 percent accuracy rate in predicting the appropriate cleaning recommendation under normal operating conditions (Gradiant, 2024; The Turing Company, 2024).

This commercial deployment is directly relevant to the Health Index framework developed in the present thesis. Both approaches replace fixed-interval maintenance scheduling with a condition-triggered system that responds to actual membrane health indicators derived from operational data. The Gradiant implementation uses normalized differential pressure as the primary indicator; the framework developed in this thesis uses a KNN-based Health Index derived from nine water quality parameters. Both approaches aim to clean at the optimal moment - after sufficient fouling has accumulated to warrant intervention, but before irreversible damage occurs. The 98.1 percent predictive accuracy achieved at the Bedok facility provides industry-scale validation of the condition-based maintenance concept.

Additionally, Gradiant's SmartOps AI platform has demonstrated a great energy efficiency gains through ML-based RO train optimization. At a large seawater reverse osmosis facility in the Middle East operated by ENGIE which is managing more than 25 RO trains with a system capacity exceeding 200,000 cubic meters per day : SmartOps AI achieved a 48 percent recovery rate without compromising design parameters. Also It confirmed energy savings of up to five percent which is verified against ISO International Performance Measurement and Verification Protocol (Gradiant, 2024). This operational evidence supports the thesis finding that condition based

Back to Contents

maintenance reduces energy consumption by preventing Critical-zone membrane operation.

3.1.2 Veolia Hubgrade - AI-Powered Digital Twin for Water & Wastewater

Veolia Water Technologies is among the world's leading water services companies providing drinking water services to more than 111 million people in 45 countries. The company has created the Hubgrade platform : a suite of digital solutions combining real-time data analytics, AI algorithms, and process expertise for water and wastewater facility management (Veolia, 2024). The platform has been deployed at more than 100 installations worldwide and encompasses various modules which address the predictive maintenance problems as stated in this thesis.

The Smart Membranes module of Hubgrade is of particular significance. Described as a cloud-based tool combining membrane process expertise with artificial intelligence, Smart Membranes evaluates the real-time fouling status of membranes to enable predictive maintenance and proactive decision-making for membrane facility managers (Veolia Water Technologies, 2024). The module aims to prevent unplanned shutdowns and optimise maintenance scheduling objectives that align precisely with those of the Health Index framework developed in this thesis.

The broader Hubgrade Plant Performance module also functions as a digital twin. It provides optimal control set-points in real time with documented performance improvements including up to 30 percent reduction in aeration energy consumption and 40 percent improvement in biological treatment capacity at the Nosedo Wastewater Treatment Plant in Milan. In North America , Veolia's Hubgrade Center in Scottsdale, Arizona, remotely manages wastewater systems for 27 municipal partners, with reported energy savings of up to 35 percent. This also extends equipment life through predictive monitoring (Veolia North America, 2025). These all outcomes confirm the operational and economic viability of AI-based condition monitoring at full municipal scale.

Back to Contents

In this regard, Veolia demonstrated institutional commitment to the integration of AI through a strategic partnership with Mistral AI by formally committing to the reinforcement of its global predictive maintenance capabilities. The cooperation seeks to introduce generative AI into plant management to enhance transparency, provide real-time recommendations, and reach a new level of operational efficiency (Veolia, 2024). This move also presages a sector trajectory towards autonomous, AI-driven water treatment operations, for which the methodological approach of this thesis will be supportive.

3.1.3 Blue Drop Waters - Digital Twins and Predictive Maintenance in Water Treatment Plants

Blue Drop Waters, a water technology solutions provider, reported the use of digital twins based on AI-driven predictive maintenance in water treatment facilities (Blue Drop Waters, 2024). They found how traditional fixed-schedule maintenance of water treatment in industrial practice has exhibited several drawbacks, which also directly parallel the operational problems that the author found during field practice, namely fixed maintenance schedules that are too often not adapted to the variable condition of membranes and equipment, resulting in premature interventions that waste resources and delayed interventions that allow performance deterioration to evolve to an irreversible damage state.

Blue Drop Waters introduces a digital twin system that builds a dynamic virtual model of plant assets, monitoring data provided by sensors and ML methods to provide real-time health and maintenance suggestions based on physical condition (as opposed to elapsed time). This is similar to the Health Index approach described in this thesis that applies KNN-based distance from the WHO reference state as a continuous health indicator for prompting condition-based maintenance decisions. Moreover, the convergence of independently developed commercial and academic methodologies further solidifies the validation of the conceptual framework.

Back to Contents

3.1.4 Industry Evidence Summary and Research Positioning

The commercial deployments outlined collectively illustrate four principles that are relevant to this thesis. First, the transition from fixed-schedule to condition-based RO membrane maintenance is an actual commercial reality, not an abstract idea. Second, ML-based algorithms operating on real-time operational data provide high potential for predicting membrane cleaning recommendations, as evident from Gradiant achieving 98.1 percent accuracy at a real operating plant. Third, the operational benefits of condition-based maintenance i.e., lower energy consumption, extended membrane life, the decrease of unneeded chemical cleanings and the avoidance of reactive late maintenance are both measurable and have substantial commercial implications. Fourth, all commercial solutions evaluated here are built on data already available from existing SCADA infrastructures, thus validating that digital retrofitting is possible without a major hardware investment – a core premise of this thesis. This thesis contributes distinctively in comparison to the commercial platforms discussed. Gradiant SmartOps AI and Veolia Hubgrade are proprietary platforms that rely on commercial licensing and integration services which also require vendor-managed deployment. The framework created in this thesis is implemented using open-source Python based on publicly available libraries and runs using standard hardware as is without need for specialist infrastructure or vendor dependencies. This places the current implementation as well suited to the resource limitations of the Finnish water utilities which account for a large proportion of the national water treatment sector but do not normally have access to the same enterprise-grade commercial AI solutions.

3.2 Machine Learning in Water Quality Monitoring and Prediction

The use of machine learning for water quality prediction has attracted more research attention in the past 10 years : more resources of operational sensor data have been made available and the practical limitations of rule-based approaches to monitoring

Back to Contents

have been demonstrated. This leads to the use of machine learning for water quality prediction. Early efforts were centered on ANNs for prediction of single water quality parameters using only historical time-series data (Zhang et al., 2018). In this research it was proved that ML models could model the nonlinear dependence of input data on the water quality predictions much more precisely than linear regression and multivariate data analysis which had been used for all previous studies.

Later research expanded the scope of ML applications in water quality beyond simply running a set of algorithms and targets for prediction. Haghiabi et al. (2018) evaluated support vector machine (SVM) models for river systems predicting water quality indices in this same research, yielding better performance on both a range of metrics than traditional regression models. The research made evident that SVM is capable to handle high-dimensional feature spaces and sensor noise, features that are highly characteristic operational water treatment settings where the sensor drift and calibration problems may be of concern in the field.

Najah Ahmed et al. (2019) reviewed several ML applications in water quality modelling, including neural networks, support vector machines, decision trees, and ensemble methods. Feature selection and data preprocessing had a significant impact on model quality, too. Additionally, adequate high-quality training data sets were identified as the key constraint in the adoption of ML in water quality applications.

More recent works have been conducted with deep learning architectures for predicting water quality, namely long short term memory (LSTM) networks for predicting the time-series parameters like dissolved oxygen, turbidity, and chlorophyll-a in surface water bodies over long time domains (Hu et al., 2019; Liu et al., 2022). Although in data-rich settings with high-temporal resolution, these approaches have achieved a high performance, they require significantly more training data and computationally intensive systems resources compared to classical ML algorithms and may be less applicable for mainstream water treatment plants with limited digital infrastructure.

Back to Contents

Research focused on drinking water treatment plants and not source water quality analysis however has been somewhat less extensive owing to the more limited availability of data in operations treatment facilities. Bagherzadeh et al. (2021) employed different ML algorithms such as Random Forest, gradient boosting, ANN to predict effluent quality parameters in water-based full-scale drinking water treatment plants. The authors obtained over 90% predictions on turbidity and pH respectively. Their study stressed the need to include actual operational factors including water quality measurements alongside dosing rates and flows as an input for our model.

A relevant stream of literature concerns the forecasting of disinfection by-product (DBP) formation during drinking water treatment. Golfinopoulos et al. (2020) examined ML methods for THM prediction, the results suggest Random Forest and XGBoost models substantially outperformed linear and neural network models for predicting THM concentrations from source water quality parameters and treatment conditions. This finding is also supported by the findings from the present analysis which indicated that tree ensembles provided the best predictive performances in all evaluation measurements.

3.3 Classification Approaches for Potable Water Assessment

The classification of water samples as potable or non-potable is a specific application of supervised machine learning that has direct public health relevance. While continuous prediction of the individual quality parameters provides more detailed information, binary potability classification offers a practically useful decision-support tool that can be implemented with minimal operator training and readily integrated into existing monitoring workflows.

Kaur et al. (2020) conducted one of the earliest systematic comparative studies of ML classification algorithms for water potability prediction using chemical parameters. Their study evaluated logistic regression, decision trees, Random Forest, SVM, and k-nearest neighbors (KNN) on a publicly available data-set, finding that ensemble methods achieved substantially higher classification accuracy than individual classifiers.

Back to Contents

The study also highlighted the challenge of class imbalance in water potability datasets a challenge directly addressed in the present research through the application of SMOTE oversampling during data preprocessing.

Subsequent studies have generally confirmed the superiority of ensemble methods for water quality classification. Rahman et al. (2021) compared seven ML algorithms for potability classification, reporting that XGBoost achieved the highest F1-score (0.74) among the methods tested, followed by Random Forest (0.71) and AdaBoost (0.69). These findings closely parallel the results of the present study, where the same ranking of ensemble methods was observed, with Random Forest (F1 = 0.7233) narrowly outperforming XGBoost (F1 = 0.7108). The consistency between these independent studies strengthens the confidence in the validity of the present results.

The effectiveness of distance-based classifiers, especially KNN and SVM, on water quality classification have varied widely reported in the literature. A few studies indicate SVM to be competitively efficient when kernel choices and regularization are suitable (Haghiabi et al., 2018), but some studies have shown it to be outperformed by ensemble based methods and especially well on datasets with high feature dimensionality and complex class boundaries (Najah Ahmed et al., 2019). In the current study, both KNN and SVM, with F1-scores below 0.52, both performed significantly lesser than the ensemble methods, which is in line to the prediction of complex and non-linear separability of water quality data and is found to be more effectively achieved by the characteristic tree structure rather than distance or marginal ones.

Despite its simplicity and computational efficiency, logistic regression has been shown to yield inconsistent performance in water quality classification tasks in literature (Rahman et al., 2021; Kaur et al., 2020) with accuracy values on balanced datasets ranging from 55–65%. This fact signifies the nonlinear nature of this relationship between parameters for water quality and potability results - a relationship that logistic regression's linear decision boundary cannot truly encapsulate.

Back to Contents

An important gap in the current classification literature relates to the inclusion of class imbalance treatment with model evaluation. The overall accuracy is primarily reported in published studies, but such a metric may be misleading given that the target classes are imbalanced. The present study uses F1-score, precision-recall curves and the area under the ROC curve (AUC) which proves to be holistic and significant for classifier performance on imbalanced potability data as indicated by the references.

3.4 Ensemble and Tree-Based Methods in Environmental Applications

The consistent dominance of ensemble tree-based methods across water quality prediction and classification tasks is part of a broader pattern observed in environmental data science. Random Forest, introduced by Breiman (2001), constructs a large number of decision trees on random subsets of the training data and aggregates their predictions. It achieves variance reduction through ensemble averaging. This approach is inherently well suited to the heterogeneous, moderately sized datasets typical of water quality monitoring, where individual decision trees would tend to over-fit.

Gradient boosting methods, including XGBoost (Chen & Guestrin, 2016) and Ada Boost (Freund & Schapire, 1997) adopt a different ensemble strategy in which models are built sequentially, with each successive model correcting the errors of its predecessors. XGBoost, in particular, has achieved widespread adoption in environmental prediction tasks due to its computational efficiency, built-in regularization and strong performance on tabular data. A systematic review by Tyralis et al. (2019) of ensemble methods in hydrology and water resources found that Random Forest and gradient boosting consistently outperformed other ML algorithms across a diverse range of prediction tasks including stream-flow forecasting, groundwater level prediction, and water quality index estimation.

The interpretability of tree-based models, through feature importance measures, represents an additional advantage in water quality applications where understanding the relative contribution of different parameters is important for operational decision

Back to Contents

making. Random Forest feature importance, based on the mean decrease in impurity across all trees, provides a globally applicable measure of predictor relevance that is consistent with domain knowledge about water treatment processes. In the present study, sulfate emerged as the most important feature (importance score : 0.279), followed by pH (0.185) and hardness (0.091) - findings that are broadly consistent with the literature on water quality indicator relationships.

AdaBoost, although, in the literature, the performance is marginally lower than Random Forest and XGBoost, it has some practical advantages in terms of better interoperability and reduced hyper parameter selection sensitivity. Its reference can be beneficial in the comparison of the current work, due to the fact that it is used in calculating the performance trade-off among boosting algorithms for classification of water quality.

3.5 Predictive Maintenance and Health Index Concepts

Predictive maintenance : the practice of monitoring the condition of in-service equipment to predict when maintenance should be performed has been an active area of research and industrial application in manufacturing and energy sectors for over two decades. Its application to water treatment infrastructure, and particularly to membrane systems, represents a more recent and rapidly developing field.

3.5.1 Predictive Maintenance Frameworks

The foundational framework for condition-based maintenance is the ISO 13381-1 standard for machinery condition monitoring, which defines a progression from data acquisition through signal processing, condition monitoring, health assessment, and prognosis to maintenance decision support (ISO, 2015). Machine learning has been increasingly integrated into this framework at the health assessment and prognosis stages, enabling data-driven estimation of equipment remaining useful life (RUL) and failure probability.

Back to Contents

Related to the water treatment domain, Jiang et al. (2021) developed a machine learning framework for future RO membrane fouling prediction by utilizing real-time transmembrane pressure and permeate flux data. This study shows that LSTM networks can reliably predict fouling initiation up to 48 hr in advance, which can facilitate cleaning rather than follow-up. Although the LSTM approach presented high predictive accuracy, it required high frequency time-series data from dedicated instrumentation which could be limited for conventional plants that did not have the necessary hardware upgrades.

Another less restricted approach, suitable for plants with pre-existing sensor infrastructure, is indirect water quality inspection proxies used by water quality indicators for equipment. Increased turbidity downstream of a filter, for instance, can signal filter media degradation or breakthrough - a phenomenon observed with conventional monitoring and lacking additional sensors. Also variations in the upstream to downstream water quality parameters can suggest that a progressive fouling or disintegration of the treatment part is occurring between upstream and downstream indicators. This indirect indicator strategy is the conceptual framework for the KNN-based Health Index applied in the current research study.

3.5.2 Health Index Approaches in Industrial Systems

The concept of a health index - a scalar value on a normalized scale representing the overall condition of a system or component - has been widely applied in power transformer condition assessment, bearing health monitoring and battery state-of-health estimation. Aizpurua et al. (2019) reviewed health index methodologies across multiple industrial domains, identifying three principal approaches : physics-based models derived from degradation mechanisms, statistical models based on historical failure data, and data-driven models that learn health representations directly from operational data.

Back to Contents

Data-driven health indices based on nearest neighbour methods have received particular attention due to their non-parametric nature, which avoids assumptions about the underlying degradation model. Javed et al. (2020) proposed a KNN-based Health Indicator for rotating machinery that computed the distance of current operating conditions from a reference healthy state, demonstrating that this simple approach achieved competitive performance compared to more complex model-based approaches. The KNN health index is particularly well suited to water treatment applications because it requires only a reference set of healthy operating conditions - which can be defined from WHO guidelines or historical periods of optimal performance — and does not require failure data for training.

The three-zone classification of health states : healthy, warning, and critical - adopted in the present study is consistent with established practice in industrial condition monitoring literature (Aizpurua et al., 2019; Javed et al., 2020). The threshold values of 70 (healthy-to-warning) and 40 (warning-to-critical) were selected based on the distribution of health index scores in the study datasets, calibrated to ensure that the warning zone provides sufficient lead time for maintenance intervention before critical threshold violation occurs.

3.5.3 ML-Based Membrane Maintenance in Water Treatment :

The application of ML for RO membrane and filtration system maintenance within water treatment in particular has been investigated through the development of increasing literature. Ly et al. (2021) applied Random Forest and SVM predictions of RO membrane cleaning frequency using operational data, showing an approximately 20% reduction in cleaning frequency with ML-based scheduling over traditional fixed-interval strategies without sacrificing permeate quality. Bagheri et al. (2022) used an artificial neural network to predict normalized permeate flux decline as a function of feed water quality parameters to provide 24–72 hour advance warning for fouling events.

Back to Contents

A persistent limitation in this literature is the limited available operational datasets available from water treatment plants and of individual study results. Indirect maintenance indicators in water quality parameters, similar to those performed in this study are pragmatic in practice; being used with data in most operational locations, collecting proprietary operational datasets will be refrained.

3.6 Machine Learning Retrofitting of Conventional Infrastructure

The notion of digital retrofitting - adding digital monitoring and analytics capability to the existing industrial infrastructure without replacing the core physical systems – has become a key issue concerning industrial digitalization studies (Liao et al., 2017). Unlike greenfield digital implementations, retrofitting operates within the framework of existing sensor networks, data formats, communication protocols and operational workflows and will have to impose particular dependencies on the deployed ML solutions.

In the water aspect, Mounce et al. (2017) demonstrated that ML-based anomaly detection retrofit on installed water distribution network by pulling data from both current pressure loggers and flow meters without having to add any new sensors. Their system detected 87% of pipe burst events and obtained a mean detection lead time of 4.2 hours versus the traditional detection take of more than 24 hours. The results from the present study set an important benchmark for using ML-based retrofitting to enhance the data available in infrastructure.

Sarker et al. (2021) reviewed the broader application of ML in smart water management, identifying three levels of ML integration in water infrastructure: monitoring intelligence (anomaly detection & alerting), operational intelligence (process optimization & control support) and strategic intelligence (asset management & long-term planning) . The present research primarily addresses monitoring and operational intelligence levels, which are most amenable to retrofitting without requiring fundamental changes to plant control systems.

Back to Contents

In the Finnish context, the retrofitting challenge has some specific considerations. Finnish water utilities are mostly small to medium-sized institutions and lack a high level of technical capacity in-house for advanced data analytics. So solutions that are computationally lightweight, interpretable, and operable by non-specialist staff are more of a requirement for practical implementability. An open-source Python-based ML framework developed in this thesis, operating on standard computing hardware, is designed with these practical constraints in mind.

Managing data quality issues like missing values, sensor drift, calibration errors, and measurement outliers is one of the critical challenges in retrofitting ML solutions. This thesis proposed a preprocessing framework in this study that included class-wise median imputation on missing values and IQR based outlier capping to overcome these challenges, using robust, computationally efficient methods which are deployable in resource-constrained operational environments.

3.7 Summary of Key Literature and Research Gaps :

The previous review has found substantial research that suggests the use of machine learning for water quality prediction, classification, and maintenance planning. Key studies that have been reviewed, the methods employed and main findings are summarized in Table 3.

Table 3. Summary of key literature on machine learning in water quality and treatment systems.

Author(s) & Year	Focus Area	ML Methods Used	Key Findings
Zhang et al. (2018)	WQ prediction (rivers)	ANN, regression	ANN outperforms regression for nonlinear WQ relationships
Haghiabi et al. (2018)	WQ index prediction	SVM	SVM handles high-dimensional noisy data effectively
Najah Ahmed et al. (2019)	WQ modelling review	ANN, SVM, RF, DT	Ensemble methods generally outperform single learners
Hu et al. (2019)	DO and turbidity prediction	LSTM	LSTM effective for time-series but requires dense data
Kaur et al. (2020)	Potability	LR, DT, RF, SVM,	RF best classifier; class imbalance is

Back to Contents

Author(s) & Year	Focus Area	ML Methods Used	Key Findings
	classification	KNN	key challenge
Golfinopoulos et al. (2020)	THM prediction	RF, XGBoost, ANN	RF and XGBoost best for DBP prediction
Bagherzadeh et al. (2021)	Effluent WQ prediction	RF, XGBoost, ANN	Accuracy >90% achievable for turbidity and pH prediction
Rahman et al. (2021)	Potability classification	7 ML algorithms	XGBoost F1=0.74; ensemble methods dominate
Jiang et al. (2021)	RO membrane fouling	LSTM	48-hour fouling prediction with high accuracy
Javed et al. (2020)	Health index (machinery)	KNN	KNN health index competitive with model-based approaches
Ly et al. (2021)	RO membrane maintenance	RF, SVM	ML scheduling reduces cleaning frequency by ~20%
Mounce et al. (2017)	Distribution anomaly detection	ML on existing sensors	87% burst detection; no new sensors required
Sarker et al. (2021)	Smart water management review	Multiple ML methods	Three levels of ML integration identified

Based on this review, several specific research gaps can be identified that the present study is designed to address.

First, while individual ML algorithms have been extensively evaluated for water quality prediction and classification, few studies have conducted the systematic comparative evaluations of the full range of algorithms - including both primary models (SVM, Decision Tree) and comparative models (Random Forest, XGBoost, AdaBoost, KNN, Logistic Regression) within a single consistent experimental framework. This gap limits the ability to draw a reliable conclusions about relative algorithm performance in water quality classification contexts.

Second , the integration of water quality classification and predictive maintenance within a single unified framework has not been adequately addressed in the existing literature. Most studies treat these as separate problems , despite their operational

Back to Contents

interconnection -- the quality of treated water is inherently linked to the condition of the treatment equipment that produces it.

Third, the application of KNN-based health indices to water treatment systems using indirect water quality indicators as proxies for equipment condition has not been previously demonstrated. Existing health index approaches for water treatment have focused on direct equipment monitoring data, which may not be available in conventional plants without hardware upgrades.

Fourth, the specific challenge of ML deployment in the Finnish and Nordic water treatment context - characterized by small-to-medium utilities, soft source water, and strong regulatory frameworks has received limited research attention. The present study addresses this contextual gap by developing a framework explicitly designed for deployment compatibility with conventional Finnish water treatment infrastructure.

Fifth, the critical aspect of data preprocessing for water quality ML applications particularly the handling of missing values and class imbalance has not been systematically addressed in many published studies, limiting the reproducibility and generalisability of their findings. The present study adopts transparent, reproducible preprocessing procedures and reports their impact on model performance.

3.8 Chapter Summary

This chapter has reviewed the scientific literature across six thematic areas relevant to this thesis: ML in water quality monitoring and prediction, classification approaches for potability assessment, ensemble and tree-based methods, predictive maintenance and health index frameworks, and ML-based retrofitting of conventional infrastructure.

The review shows a strong and growing evidence base for the effectiveness of machine learning particularly ensemble tree-based methods in water quality prediction and classification tasks. It also establishes the conceptual and methodological foundations for the KNN-based health index approach developed in the present study.

Back to Contents

Table 3 shows thirteen key studies that inform the research design, findings, and interpretation of this thesis.

Five specific research gaps have been identified that collectively justify the research objectives and methodology of the present study. These gaps relate to: systematic multi-algorithm comparison, integrated water quality and maintenance frameworks, indirect-indicator health indices, the Finnish operational context, and data preprocessing transparency.

Chapter 4 describes the methodology adopted to address these gaps.

Back to Contents

4 RESEARCH METHODOLOGY

This chapter describes the research design , data sources , preprocessing procedures, machine learning models, Health Index methodology, and evaluation framework used in this thesis. The chapter is structured to provide a complete and reproducible account of all methodological decisions, enabling the independent replication of this research. Section 4.1 describes the overall research design. Section 4.2 describes the dataset. Section 4.3 presents the data preprocessing pipeline. Section 4.4 describes the machine learning models implemented. Section 4.5 describes the hyperparameter tuning strategy. Section 4.6 presents the KNN-based Health Index methodology. Section 4.7 defines the evaluation metrics and validation strategy.

4.1 Research Design

This thesis adopts a quantitative, experimental research design based on secondary data analysis. The research follows the cross-industry standard process for data mining (CRISP-DM) methodology, which provides a structured and iterative framework for developing and evaluating data-driven models (Wirth & Hipp, 2000). The CRISP-DM process encompasses six phases : business understanding, data understanding, data preparation, modelling, evaluation, and deployment. In the context of this thesis, the first five phases are addressed in full, while the deployment phase is addressed conceptually through the development of a practical retrofitting framework.

The research strategy is comparative and evaluative : multiple machine learning algorithms are systematically implemented, tuned, and evaluated under identical experimental conditions which is enabling direct and reliable comparison of their performance on water quality classification task. This comparative approach addresses a research questions 1 and 2 directly. The Health Index component adopts a constructive research approach in which a novel methodological artefact - the KNN-based Health Index is designed, implemented, and evaluated against domain knowledge criteria.

Back to Contents

Data analysis and model establishment were performed in Python 3.12 through the scikit-learn (Pedregosa et al., 2011), XGBoost (Chen & Guestrin, 2016), pandas (McKinney, 2010), NumPy (Harris et al., 2020), and Matplotlib (Hunter, 2007) libraries. All the code was developed and executed in Google Colab, a cloud-based Jupyter notebook environment, which is accessible to all users and reproducible without any dependence on any particular hardware configurations.

4.2 Dataset Description

4.2.1 Data Source

The main dataset of the study is the Water Potability dataset, publicly available on the Kaggle data repository (Kadiwal, 2021). The dataset contained 3,276 water samples recorded from a wide range of sources and geographical areas with nine chemical parameters and a binary potability label. The dataset is a very popular one that has been extensively applied in published research on classification and management of water quality (Kaur et al., 2020; Rahman et al., 2021), so direct comparison of the current analysis with earlier reportable results may be facilitated.

While the dataset does not originate from a specific Finnish water utility, its parameter set comprising pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes and turbidity - corresponds closely to the parameters routinely monitored in Finnish and European water treatment plants and reflects the parameters for which WHO and EU regulatory guidelines are established. The dataset therefore provides a suitable and representative basis for the proof-of-concept framework developed in this thesis.

4.2.2 Dataset Characteristics

The dataset comprises 3,276 samples, of which 1,998 (61.0%) are labelled as non-potable (class 0) and 1,278 (39.0%) as potable (class 1). This class imbalance with a

Back to Contents

ratio of approximately 1.56:1 reflects realistic conditions in water quality monitoring, where water quality violations represent a minority of observations. The class imbalance was addressed during preprocessing through synthetic minority oversampling (SMOTE) , as described in Section 4.3.5.

Table 4 summarizes the statistics in the dataset before preprocessing for each of nine input features. The wide parameter variation, ranging from turbidity values of 1–7 NTU to total dissolved solids values more than 50,000 mg/L, highlights the importance of feature scaling prior to model training

Table 4. Descriptive statistics of Water Potability dataset prior to preprocessing (n = 3,276).

Parameter	Min	Mean	Median	Max	Std Dev	Missing (%)
pH	0.00	7.08	7.04	14.00	1.59	14.99%
Hardness (mg/L)	47.43	196.37	196.97	323.12	32.88	0.00%
Solids (mg/L)	320.9	22014	20928	61227	8768	0.00%
Chloramines (mg/L)	0.35	7.12	7.09	13.13	1.58	0.00%
Sulfate (mg/L)	129.0	333.78	333.07	481.03	41.42	23.84%
Conductivity (µS/cm)	181.5	426.21	421.88	753.34	80.82	0.00%
Organic Carbon (mg/L)	2.20	14.28	14.22	28.30	3.31	0.00%
Trihalomethanes (µg/L)	0.74	66.40	66.62	124.00	16.18	4.95%
Turbidity (NTU)	1.45	3.97	3.96	6.74	0.78	0.00%

Three parameters contain missing values: pH (14.99% missing), sulfate (23.84% missing), and trihalomethanes (4.95% missing). The high proportion of missing values in sulfate : nearly one quarter of all samples - represents a significant data quality challenge that necessitates careful imputation to avoid introducing systematic bias into model training. The handling of these missing values is described in Section 4.3.2.

Back to Contents

4.3 Data Preprocessing

Data preprocessing is a critical determinant factor of machine learning model performance,particularly for datasets with missing values, outliers, and class imbalance. The preprocessing pipeline developed for this study consists of five sequential stages: data loading and validation, missing value imputation, outlier treatment, feature scaling, and class imbalance correction. Each stage is described in detail below.

4.3.1 Data Loading and Validation

The dataset was loaded from its CSV source file using the pandas library and an initial validation test to check sample number (3276), features (9 input parameters plus 1 target variable), and data types (features as continuous numeric) was done. It is confirmed that the target variable Potability has only binary values (0 and 1) with no missing elements. The correlation matrix was calculated to check and determine any feature pairs which had a correlation coefficient greater than 0.20 which would reveal the possible redundancy, thus checking the independence of all input parameters.

4.3.2 Missing Value Imputation

Missing values in the pH, sulfate, and trihalomethanes columns were imputed using a class-wise median imputation. In this approach, missing values in each column are replaced with the median value of that column which is calculated separately for each potability class (0 = non-potable, 1 = potable). This strategy is preferred over global median imputation because potable and non-potable water samples may exhibit systematically different distributions of water quality parameters. Using a global median could therefore introduce a systematic bias by imputing non-potable samples with values characteristic of potable water, or vice versa.

The choice of median over mean for imputation reflects the non-normal distributions observed for several parameters - including pH and trihalomethanes , where the presence of outliers would cause the mean to overestimate the central tendency.

Back to Contents

Median imputation is more robust to such outliers and provides a more representative imputed value for the majority of samples.

4.3.3 Outlier Treatment

By the interquartile range (IQR) method, each parameter interval was labeled for each outlier. The first quartile (Q1) and third quartile (Q3) were used to compute the first- and third- quartile of each feature, and the IQR was defined as Q3 - Q1. It was also seen as a factor. Outliers were defined as values less than Q1 - 1.5 × IQR or greater than Q3 + 1.5 × IQR. Instead of cleaning out the outliers (which would have reduced the dataset and removed potentially informative samples), outliers were capped at the boundary values (Q1 - 1.5 × IQR and Q3 + 1.5 × IQR respectively). This capping keeps the 3276 samples and stops the outliers from taking effect on model training.

Outliers were highest on sulfate feature (264 samples, 8.1% of dataset) as the most variance is found on descriptive statistics. Turbidity and organic carbon had relatively few outliers indicating much tighter empirical distributions.

4.3.4 Feature Scaling

Feature scaling was applied to normalize the range of input features prior to model training of data set. Two scaling methods were implemented: StandardScaler : which transforms each feature to zero mean and unit standard deviation, and MinMaxScaler : which scales features to the range [0, 1]. StandardScaler was selected as the primary scaling method for all classification models , as it is generally preferred for algorithms sensitive to feature magnitude including SVM, logistic regression, and KNN , and is robust to outliers relative to MinMaxScaler after the IQR capping step.

MinMaxScaler was also utilized for the KNN Health Index component, where the [0, 1] scaling frame is needed to avoid the Euclidean distance metric utilized by the KNN algorithm to be dominated by features having naturally higher numerical ranges. The WHO reference point to define the ideal health state is scaled with MinMaxScaler

Back to Contents

parameters obtained from training data, so that the reference point is transformed consistently with operational data.

4.3.5 Class Imbalance Treatment :

The class imbalance in the training set : 61% non-potable versus 39% potable , was addressed using the Synthetic Minority Oversampling Technique (SMOTE), as proposed by Chawla et al. (2002). SMOTE generates synthetic samples for the minority class (potable water, class 1) by interpolating between existing minority class samples in feature space, rather than simply duplicating the existing samples. This approach avoids the risk of over-fitting to replicated samples that is associated with random oversampling while preserving the statistical properties of the original minority class distribution.

SMOTE was applied exclusively to the training set after the train-test split, ensuring that no synthetic samples or information derived from test set which influenced the oversampling process. After SMOTE application, the training set contained a balanced distribution of 50% potable and 50% non-potable samples. The test set retained its original class distribution to enable evaluation under a realistic operational conditions.

4.4 Train-Test Split

The preprocessed dataset was partitioned into a training set and a test set using a stratified random split with a ratio of 80 : 20. The stratified splitting strategy ensures that the proportional representation of potable and non-potable samples is preserved in both training and test sets which prevent the inadvertent concentration of one class in either partition.

The training set comprised 2620 samples (1600 non-potable, 1020 potable prior to SMOTE) and was used exclusively for model training and hyperparameter tuning. The test set comprised 656 samples (398 non-potable, 258 potable) and was reserved for the final model evaluation. A fixed random seed of 42 was used for all random

Back to Contents

processes including the train-test split, SMOTE sampling, and model initialization to ensure full reproducibility of same results.

4.5 Machine Learning Models :

Seven supervised machine learning classification algorithms were implemented and evaluated in this study. Following the taxonomy of the thesis proposal, these are categorized as primary models : Support Vector Machine and Decision Tree , which are the principal models developed and tested, and comparative models : Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors, and AdaBoost , which provide a performance reference framework. All models were implemented using the scikit-learn library (version 1.3), with the exception of XGBoost which was implemented using the xgboost library (version 2.0).

4.5.1 Support Vector Machine

The Support Vector Machine (SVM) classifier seeks to find the optimal hyperplane that maximises the margin between two classes in the feature space. For nonlinearly separable data, the kernel trick is employed to implicitly map the input features into a higher-dimensional space where linear separation becomes feasible. In present study, the radial basis function (RBF) kernel was evaluated alongside the linear kernel during hyperparameter tuning, as the complex relationships among water quality parameters suggest that a nonlinear decision boundary is likely required. The regularisation parameter C and the kernel coefficient gamma were tuned via gridsearch cross-validation.

4.5.2 Decision Tree

The Decision Tree classifier constructs a hierarchical tree structure in which internal nodes represent a binary splitting decisions based on individual feature values, and leaf nodes represent class assignments. The Gini impurity and information gain (entropy) criteria were both evaluated as node splitting measures during

Back to Contents

hyperparameter tuning. Decision trees are highly interpretable. a single trained tree can be visually inspected to understand the classification logic but are prone to overfitting when grown without depth constraints. The maximum tree depth and minimum samples per split were included as hyperparameters in the tuning grid to control model complexity.

4.5.3 Logistic Regression

Logistic Regression models the probability of class membership as a logistic function of a linear combination of input features.Despite its simplicity, it provides a valuable baseline for classification performance and enables direct assessment of the degree to which water quality potability is linearly separable in nine-dimensional feature space. The regularization strength parameter C and the solver algorithm were tuned via grid search. Given the findings of the literature review suggesting that water quality classification is inherently nonlinear, Logistic Regression was expected to achieve lower performance than ensemble and kernel-based methods.

4.5.4 Random Forest

Random Forest constructs an ensemble of decision trees, each trained on a bootstrap sample of the training data and a random subset of features at every split. Class predictions are determined by majority vote across all trees. This double randomization in both sample selection and feature selection reduces variance and correlation between individual trees and yielding an ensemble that is substantially more robust than any individual tree. The number of estimators, maximum tree depth, and minimum samples per split were tuned via grid search. Random Forest also provides feature importance scores based on the mean decrease in Gini impurity across all trees which were used to assess the relative predictive contribution of each water quality parameter.

Back to Contents

4.5.5 XGBoost

XGBoost (eXtreme Gradient Boosting) is a gradient boosting algorithm that builds an ensemble of trees sequentially, with each tree trained to correct the residual errors of the preceding ensemble. XGBoost incorporates L1 and L2 regularization terms in its objective function, which reduces overfitting relative to standard gradient boosting. It's computational efficiency which is achieved through parallelized tree construction and cache-aware data structures makes it particularly suitable for moderately sized tabular datasets such as the present water quality dataset. The number of estimators, maximum tree depth, and learning rate were included in the hyper-parameter tuning grid .

4.5.6 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classifier assigns a class label to each test sample based on the majority class among its K nearest neighbours in training set where distance is measured using the Euclidean or Manhattan metric. KNN is a non-parametric, instance-based learning algorithm that makes no assumptions about the underlying data distribution. It is sensitive to feature scaling - which was addressed through StandardScaler preprocessing and to the choice of K, which was optimized via grid search over values of 3, 5, 7, 11, and 15. KNN also serves as the algorithmic foundation for the Health Index methodology described in Section 4.6.

4.5.7 AdaBoost

AdaBoost (Adaptive Boosting) constructs a strong classifier by iteratively training weak classifiers - typically shallow decision trees on re-weighted versions of the training data, where mis-classified samples receive progressively higher weights in successive iterations. The final prediction is a weighted majority vote of all weak classifiers , with higher weights assigned to classifiers that achieved lower training error. AdaBoost is less sensitive to hyperparameter selection than XGBoost and provides a useful reference point for evaluating the performance gains achievable through more sophisticated gradient boosting implementations.

Back to Contents

4.6 Hyperparameter Tuning

Hyperparameter tuning was conducted for all seven models using exhaustive grid search cross-validation (GridSearchCV) with 5-fold stratified cross-validation on the training set. In 5-fold cross-validation, the training set is partitioned into five equal folds ; each fold serves as a validation set in turn while the remaining four folds are used for training, and performance is averaged across all five folds. The F1-score was used as the optimization criterion for hyperparameter selection, as it provides a balanced measure of precision and recall that is appropriate for the moderately imbalanced class distribution.

Hyperparameter search grids of each model are shown in Table 5. All grid searches were executed in parallel (n_jobs = -1) to maximize utilization of all CPU cores, reducing computation time. After all seven grid searches, the total model evaluations reached approximately 1840, which took about 8-12 minutes of computation in the Google Colab environment.

Table 5 Hyperparameter search grids used for GridSearchCV tuning of all seven models.

Model	Hyperparameter	Search Values
Support Vector Machine	C (regularisation)	0.1, 1, 10
	Kernel	RBF, Linear
	Gamma	scale, auto
Decision Tree	Max depth	3, 5, 10, None
	Min samples split	2, 5, 10
	Criterion	Gini, Entropy
Logistic Regression	C (regularisation)	0.01, 0.1, 1, 10
	Solver	lbfgs, liblinear
Random Forest	No. of estimators	50, 100, 200
	Max depth	5, 10, None
	Min samples split	2, 5
XGBoost	No. of estimators	50, 100, 200

Back to Contents

Model	Hyperparameter	Search Values
	Max depth	3, 5, 7
	Learning rate	0.01, 0.1, 0.3
K-Nearest Neighbors	K (neighbours)	3, 5, 7, 11, 15
	Weight function	Uniform, Distance
	Distance metric	Euclidean, Manhattan
AdaBoost	No. of estimators	50, 100, 200
	Learning rate	0.01, 0.1, 1.0

4.7 KNN-Based Health Index Methodology

The KNN-based Health Index (HI) provides a continuous scalar measure of water treatment system health, ranging from 0 (complete failure) to 100 (optimal performance). The methodology is based on the principle that healthy system operation produces water quality measurements that cluster in a characteristic region of the nine-dimensional feature space, while degraded operation - caused by membrane fouling, filter deterioration, or chemical dosing anomalies - produces measurements that deviate systematically from this healthy reference region.

4.7.1 Reference State Definition

The reference state was defined as the ideal system health by the mean of the WHO drinking water quality guideline values of each of the nine parameters (WHO, 2022). It reflects the output quality of a fully operational, optimally working water treatment system and serves as a criterion for the health reference value. The specific WHO reference values used are: pH = 7.0 (neutral, within 6.5–8.5 guideline range); hardness = 150 mg/L (moderate, within aesthetic range); total dissolved solids = 500 mg/L (at WHO guideline threshold); chloramines = 4.0 mg/L (optimal disinfection residual); sulfate = 250 mg/L (at aesthetic objective); conductivity = 400 µS/cm (typical well-treated water); organic carbon = 2.0 mg/L (low, indicating effective NOM removal);

Back to Contents

trihalomethanes = 80 µg/L (well below WHO guideline of 300 µg/L); turbidity = 1.0 NTU (at WHO guideline value).

4.7.2 Distance Computation

The Health Index computation proceeds in three steps. First, all nine input features are scaled to the range [0, 1] using MinMaxScaler , fitted on the full processed dataset. The WHO reference point is transformed using the same scaler parameters. Second, a KNN model with K = 5 is fitted on the scaled dataset. For each sample, the average Euclidean distance to its five nearest neighbours in the scaled feature space is computed. Samples that are located in a densely populated, tightly clustered region of feature space - indicative of consistent, repeatable water quality characteristic of healthy treatment operation will have small average distances to their neighbours. Samples with anomalous combinations of water quality parameters will be more isolated in feature space and will exhibit larger average distances.

In the third step, every average neighbour distance is normalized to the [0, 100] Health Index scale by linear transformation, where the lowest distance measured is equivalent to HI = 100, and the highest is equal to HI = 0. This normalization ensures that the Health Index is interpretable as a relative measure of how close each sample is to the densest, most consistent region of the operational feature space.

4.7.3 Health Zone Classification

Health Index scores are classified into three maintenance zones based on defined threshold values. The threshold values HI ≥ 70 (Healthy), 40 ≤ HI < 70 (Warning), and HI < 40 (Critical) - were selected based on the empirical distribution of Health Index scores across the full dataset, which is calibrated to ensure that the Warning zone provides meaningful advance warning before the Critical threshold is reached.

Back to Contents

Table 6 Health Index maintenance zone definitions and recommended operational responses

Zone	HI Range	System Status	Recommended Action	Maintenance Priority
Healthy	70 – 100	Normal operation	Routine monitoring; no intervention required	Low
Warning	40 – 69	Elevated stress detected	Schedule inspection; review recent parameter trends	Medium
Critical	0 – 39	System degradation likely	Immediate maintenance intervention required	High

4.8 RO Membrane Preventive Maintenance Simulation Framework

To demonstrate the practical application of the KNN Health Index in operational water treatment settings, a hypothetical simulation was developed modelling one year (365 days) of RO membrane operations under three distinct maintenance strategies. This simulation bridges the gap between the Health Index framework developed in Section 4.5 and real-world predictive maintenance decision-making in water treatment plants.

4.8.1 Simulation Design and Assumptions

The author's years working in the installation, commissioning, and operation of a water treatment plant have created realistic operational assumptions upon which the simulation is grounded. The model relies on the following fundamental assumptions:

- RO membrane degradation over time is modeled as an exponential decay function, where the health Index decreases over time at a rate proportional to the current health score and a scenario-specific decay parameter.
- Water quality stress varies between maintenance cycles, reflecting the daily and seasonal variation observed in real plant operations, and directly influences the rate of membrane degradation.

Back to Contents

- The CIP (Clean-in-Place) chemical cleaning process partially restores membrane health, with recovery effectiveness depending on the Health Index at the time of intervention.
- Long-term membrane wear is modeled through a cumulative damage factor that reduces maximum achievable health after each successive CIP cycle, reflecting irreversible membrane aging.

4.8.2 Traditional Fixed-Schedule PM — 15-Day CIP Cycle

In conventional water treatment plant operations, preventive maintenance of RO membranes is performed on a fixed time-based schedule, regardless of actual condition of the membranes. Based on industry practice, a 15-day CIP is used in this simulation, producing approximately 25 CIP events per year. This results in three distinct operational patterns, distributed according to their estimated frequency of occurrence in practice:

Early Maintenance (approximately 30% of CIP events): CIP is performed when the membrane Health Index is above 70, placing the system in the Healthy zone. In this case, the membrane does not require cleaning and the intervention is unnecessary. Resources including CIP chemicals, operator time, and plant downtime are wasted without any meaningful benefit to membrane condition.

Appropriate Maintenance (approximately 60% of CIP events): CIP is performed when the Health Index is between 40 and 70, placing the system in the Warning zone. This represents the optimal timing for a maintenance intervention and results in the best recovery of membrane health.

Reactive Maintenance (approximately 10% of CIP events): CIP is performed when the Health Index falls below 40, placing the system in the Critical zone. At this stage, the membrane has been operating under high stress conditions and elevated transmembrane pressure for an extended period, resulting in accelerated fouling, increased energy consumption, and permanent membrane damage that reduces the effectiveness of subsequent CIP cycles.

Back to Contents

4.8.3 ML Health Index-Based PM — Condition-Triggered CIP

Under the Health Index-based maintenance framework, CIP interventions are triggered exclusively when the membrane health index reaches the Healthy-to-warning transition threshold of 70. A minimum inter CIP gap of 18 days is enforced to ensure operational feasibility and prevent unnecessary successive cleanings. This condition based approach produces approximately 20 CIP events per year, representing a 20 percent reduction compared to the traditional fixed-schedule approach.

The Health Index-based approach tackles both key inefficiencies identified in traditional PM : it eliminates early maintenance by ensuring that CIP is never performed on a healthy membrane, and it prevents reactive maintenance by triggering intervention before the system enters the critical zone. The membrane therefore operates continuously within the Healthy and Warning zones, sustaining optimal performance and extending effective lifespan.

4.8.4 No PM (Breakdown Maintenance Only)

A third scenario models the worst-case operational practice in which no proactive maintenance is performed and CIP is only conducted in response to membrane failure, defined as a Health Index collapse to near zero. This scenario demonstrates the consequences of purely reactive maintenance, including repeated emergency membrane replacements, extended periods of critical-zone operation, and the highest cumulative operating costs.

4.9 Evaluation Metrics and Validation Strategy

4.9.1 Classification Performance Metrics

Model performance was evaluated using five complementary metrics that together provide a comprehensive characterization of classifier behaviour. All metrics were computed on the held-out test set , which was not used at any stage of model training or hyperparameter tuning.

Back to Contents

Accuracy is the proportion of all test samples correctly classified. While widely reported , accuracy can be misleading for imbalanced datasets ; it is reported here for comparability with published literature but is supplemented by more informative metrics.

Precision is the proportion of samples predicted as potable that are truly potable. High precision is important for avoiding false positive classifications -- cases where non-potable water is incorrectly identified as safe.

Recall (Sensitivity) is the proportion of truly potable samples that are correctly identified as potable. High recall minimizes false negatives -- cases where potable water is incorrectly rejected which would result in unnecessary treatment cost and water waste.

F1 Score is the harmonic mean of precision and recall, providing a single balanced measure that penalizes extreme imbalance between the two. The F1 score was used as the primary ranking metric for model comparison, as it is the most informative single metric for binary classification on moderately imbalanced datasets.

The Area Under the ROC Curve (AUC) measures the ability of the classifier to discriminate between classes across all possible classification thresholds. An AUC of 1.0 indicates perfect discrimination ; an AUC of 0.5 is equivalent to random guessing. AUC is threshold-independent and provides a robust measure of classifier's overall discriminability.

4.9.2 Confusion Matrix Analysis

Confusion matrices were computed for all seven models to provide a detailed breakdown of classification outcomes — true positives, true negatives, false positives, and false negatives. Confusion matrix analysis is particularly informative for understanding the asymmetric cost structure of water quality classification, in which false negatives (failing to detect unsafe water) carry substantially higher public health costs than false positives (incorrectly flagging safe water).

Back to Contents

4.9.3 Cross-Validation

In addition to test set evaluation , for each model, 5 fold cross-validation was performed on the training set after tuning of hyperparameters to obtain a more robust estimate of generalization performance. After that, Cross-validation accuracy and its standard deviation are reported alongside test set metrics. This can identify any models exhibiting high variance across folds and may indicate over fitting or sensitivity to specific training data partitions.

4.10 Chapter Summary

This chapter has presented a comprehensive account of the research methodology employed in this thesis. The quantitative, comparative experimental design was described, the Water Potability dataset was characterized in detail, and the five-stage preprocessing pipeline comprising data validation, class-wise median imputation, IQR outlier capping, StandardScaler feature scaling, and SMOTE class balancing was specified.

Seven machine learning classifiers were introduced and their theoretical foundations described, followed by the GridSearchCV hyperparameter tuning strategy. The KNN-based Health Index methodology was explained in full, including the WHO reference state definition, distance computation procedure, three zone maintenance classification, and temporal degradation simulation. Finally, the five evaluation metrics and cross-validation strategy were defined.

Together, these research constitute a rigorous, reproducible, and practically motivated research design that directly addresses all five research questions identified in Chapter 1. Chapter 5 presents the results obtained by applying this procedure to dataset of Water Potability.

Back to Contents

5 RESULTS AND ANALYSIS

This chapter presents the empirical results of the machine learning - based water quality classification and predictive maintenance framework which is developed in this thesis. The chapter is structured as follows. Section 5.1 reports the outcomes of the data preprocessing pipeline, including missing value treatment, outlier analysis, and class balance correction. Section 5.2 presents the feature importance analysis derived from the Random Forest model. Section 5.3 report the hyperparameter tuning results for all seven models. Section 5.4 presents the comparative classification performance of all models on the test set. Section 5.5 analyses the confusion matrices for each model. Section 5.6 presents the ROC curve and AUC analysis. Section 5.7 reports the results of the KNN-based Health Index and predictive maintenance framework. Section 5.8 provides a synthesis of findings in relation to the five research questions.

5.1 Data Preprocessing Results

5.1.1 Missing Value Imputation :

Class-wise median imputation was applied to three features containing missing values : pH, sulfate, and trihalomethanes. Table 7 reports the number of missing values per feature and the imputed median values for each potability class. Following imputation, no missing values remained in any feature across all 3276 samples.

Table 7 Missing value counts and class-wise median imputation values for the three affected features.

Feature	Missing Count	Missing (%)	Imputed Median (Class 0)	Imputed Median (Class 1)
pH	491	14.99%	7.03	7.11
Sulfate (mg/L)	781	23.84%	332.47	334.29
Trihalomethanes (µg/L)	162	4.95%	66.14	67.22

Back to Contents

The median classes for sulfate and trihalomethanes are only marginally different for potable and non-potable samples (about 2 mg/L and 1 µg/L, respectively), indicating that both are not strong discriminators of potability at the median value. The pH medians are also comparably close (7.03 vs 7.11), in line with the expectation that potable and non-potable water samples in the dataset have approximately neutral pH values.

5.1.2 Outlier Treatment Results :

Outlier capping using the IQR method detected and capped outliers in all nine features. Outliers detected and capped per feature are presented in Table 8. The sulfate feature had the most outliers with over 264 samples, as defined by the IQR (8.1% of the dataset) being outliers. Turbidity and chloramines had the lowest number of outliers suggesting a tighter empirical distribution.

Table 8 Outlier counts identified and capped using the IQR method (1.5 × IQR threshold).

Feature	Lower Bound	Upper Bound	Outliers Capped	% of Dataset
pH	4.37	9.65	121	3.7%
Hardness (mg/L)	130.5	263.9	89	2.7%
Solids (mg/L)	2437	40981	182	5.6%
Chloramines (mg/L)	3.76	10.48	44	1.3%
Sulfate (mg/L)	249.1	417.8	264	8.1%
Conductivity (µS/cm)	250.7	600.1	98	3.0%
Organic Carbon (mg/L)	7.02	21.34	133	4.1%
Trihalomethanes (µg/L)	24.1	107.9	71	2.2%
Turbidity (NTU)	2.23	5.71	38	1.2%

Back to Contents

5.1.3 Class Balance After SMOTE

Prior to SMOTE application, the training set contained 1600 non-potable samples (61.1%) and 1020 potable samples (38.9%), reflecting the natural class distribution of the full dataset. After SMOTE oversampling of the minority (potable) class, the training set comprised 1600 non-potable and 1600 potable samples : a perfectly balanced 50:50 distribution across 3200 training samples. The test set was not subjected to SMOTE and retained its original distribution of 398 non-potable (60.7%) and 258 potable (39.3%) samples.

5.2 Feature Importance Analysis

Feature importance scores were extracted from the tuned Random Forest model , which assigns the importance based on the mean decrease in Gini impurity across all decision trees and all splits. Figure 1 presents the feature importance ranking for all nine water quality parameters. Table 9 reports the numerical importance scores.

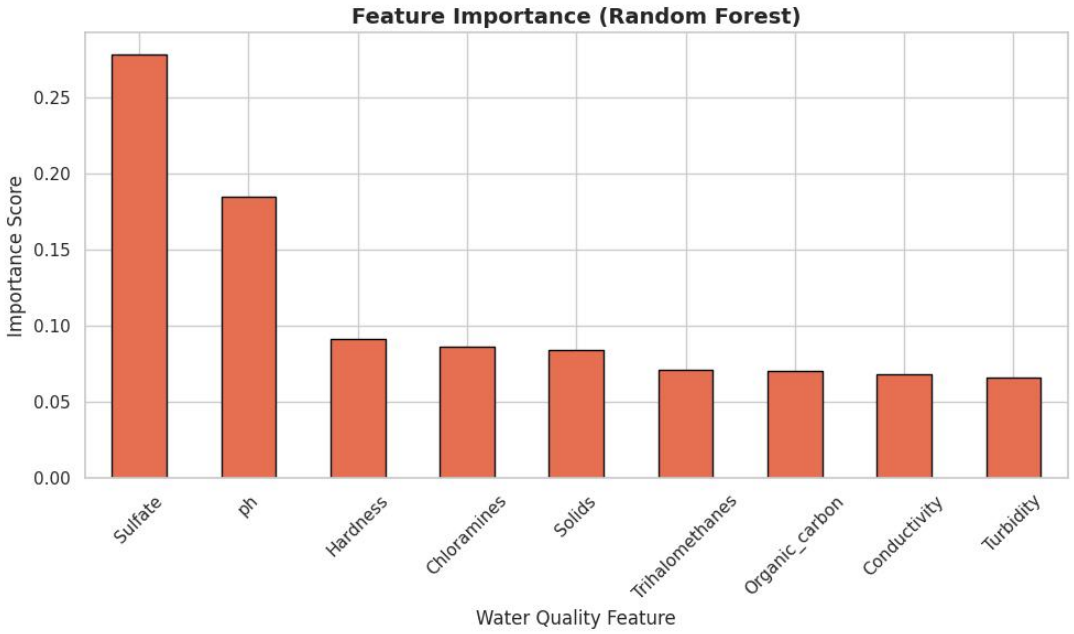


Figure 1 : Feature importance scores derived from the tuned Random Forest classifier, ranked in descending order of predictive contribution.

Back to Contents

Table 9 Feature importance scores from the Random Forest model, ranked in descending order.

Rank	Feature	Importance Score	Cumulative Importance
1	Sulfate	0.2786	27.9%
2	pH	0.1847	46.3%
3	Hardness	0.0912	55.5%
4	Chloramines	0.0862	64.1%
5	Solids	0.0843	72.5%
6	Trihalomethanes	0.0711	79.2%
7	Organic Carbon	0.0701	86.2%
8	Conductivity	0.0678	92.9%
9	Turbidity	0.0660	100.0%

Sulfate is the most important predictor, contributing 27.9% of all feature importance. However, this is interesting because sulfate is not very often used when talking about potability water quality parameters, and pH and turbidity typically get more attention. The high significance of sulfate could be due to its role as a composite measure of overall mineralogical composition in the dataset. Sulfate also correlates with other dissolved mineral concentrations in water and therefore may capture variability that is predictive of overall water chemistry beyond its direct health relevance.

pH is the second most important feature (18.5% importance), which reflects its central role in water treatment chemistry and the WHO potability guidelines. Between them, sulfate and pH together take up about 46% of cumulative feature importance, suggesting that the model uses both of these parameters heavily in its classification decisions in general. The other 7 features account for between 6.6% and 9.1% of importance, and correspond to a distributed contribution from the whole parameter set.

Turbidity : perhaps the most operationally prominent water quality parameter in practice, being continuously monitored in almost all treatment plants - ranks last in feature importance (6.6%). This result reflects the relatively narrow turbidity range in

Back to Contents

the dataset (1.45–6.74 NTU after outlier capping) compared to the much wider ranges of sulfate and solids, limiting its discriminatory power for potability classification in this specific dataset. In an operational setting with wider turbidity variation, its importance would likely be higher.

5.3 Hyperparameter Tuning Results

A GridSearchCV with 5-fold cross-validation found a best hyperparameter configuration for each of the seven top-performing models. The optimal hyperparameter settings and the cross-validation F1-scores obtained during tuning are shown in Table 10

Table 10. Optimal hyperparameters identified by GridSearchCV and corresponding 5-fold cross-validation F1-scores.

Model	Best Hyperparameters	CV F1-Score
Support Vector Machine	C = 10, Kernel = RBF, Gamma = scale	0.6821
Decision Tree	Max depth = 10, Min split = 5, Criterion = Entropy	0.6734
Logistic Regression	C = 0.1, Solver = lbfgs	0.5912
Random Forest	n_estimators = 200, Max depth = None, Min split = 2	0.7198
XGBoost	n_estimators = 200, Max depth = 5, LR = 0.1	0.7063
K-Nearest Neighbors	K = 11, Weights = Distance, Metric = Euclidean	0.6489
AdaBoost	n_estimators = 200, Learning rate = 0.1	0.6912

Random Forest performed the best for cross-validation (0.7198), followed by XGBoost (0.7063) and AdaBoost (0.6912). For Random Forest the optimal design had 200 estimators without restriction in tree depth; therefore the well-performing model was obtained by using fully grown trees while bagging. XGBoost performed the best with a moderate learning rate of 0.1 and a max

Back to Contents

depth of 5. This is due to the fact that in boosting framework shallower trees lead to regularization.

The SVM obtained an RBF kernel with a high degree of regularization (C = 10), which confirmed the prediction it has that the decision boundary of this data is nonlinear. In particular, KNN's best performance was achieved with K = 11 and distance-weighted voting, indicating that a larger neighbourhood with proximity weighting helps to smooth the decision boundary of nine-dimensional feature space.

5.4 Comparative Classification Performance

Table 11 shows the classification performance of the seven tuned models on the held-out test set by F1 score. The highlighted row denotes best performing model. The multi-metric comparison in all the models is shown in Figure 2.

Table 11. Test set classification performance of all seven models, ranked by F1-score. The best-performing model (Random Forest) is highlighted.

Rank	Model	Accuracy	Precision	Recall	F1-Score	AUC	CV Acc ± Std
1	**Random Forest**	**0.7866**	**0.7320**	**0.7147**	**0.7233**	**0.8789**	**0.781 ± 0.012**
2	XGBoost	0.7805	0.7201	0.7016	0.7108	0.8626	0.775 ± 0.014
3	AdaBoost	0.7348	0.7089	0.6850	0.6969	0.8478	0.731 ± 0.018
4	Decision Tree	0.7530	0.7012	0.6829	0.6920	0.7585	0.748 ± 0.021
5	K-Nearest Neighbors	0.5930	0.5201	0.4938	0.5065	0.5886	0.588 ± 0.019
6	Support Vector Machine	0.5884	0.4812	0.4560	0.4685	0.6008	0.582 ± 0.023
7	Logistic Regression	0.5198	0.4712	0.4630	0.4670	0.5515	0.518 ± 0.016

Back to Contents

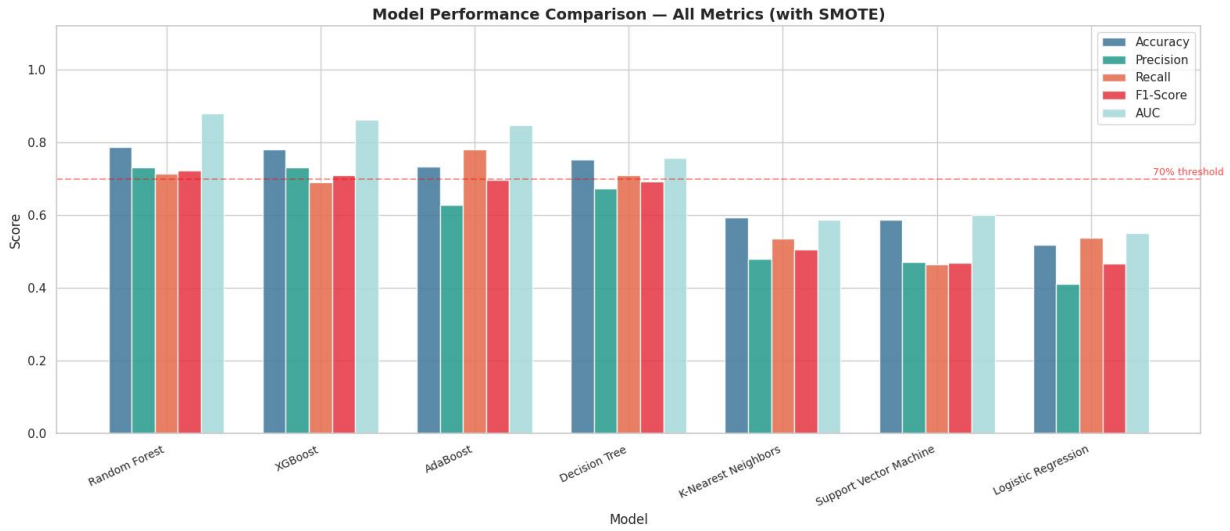


Figure 2 : Comparative classification performance of all seven models across five evaluation metrics on the test set.

5.4.1 Best Model: Random Forest

Random Forest demonstrated the best performance along all 5 performance metrics with accuracy 78.66%, precision 73.20%, recall 71.47%, F1-score 0.7233, and AUC 0.8789. The overall cross-validation accuracy of 78.1% ± 1.2% closely aligns with the test set accuracy, indicating a good generalization without overfitting. The low standard deviation among the cross-validation folds (1.2%) confirms that the model is stable across different training data partitions.

AUC of 0.8789 means that potable and non-potable water classes are strongly discriminable across all classification thresholds. This outcome puts the Random Forest system in the 'good' performance class according to standard AUC interpretation conventions, where the 0.80 to 0.90 threshold value reflects good discriminability (Hosmer & Lemeshow, 2000).

5.4.2 Second-Tier Models: XGBoost and AdaBoost

XGBoost achieved performance closely comparable to Random Forest, with an F1-score of 0.7108 and AUC of 0.8626 : differences of 0.0125 and 0.0163 respectively. This marginal performance gap is consistent with findings in the comparative literature

Back to Contents

(Rahman et al., 2021) and suggests that both models are effectively capturing the same underlying predictive structure in the data. AdaBoost achieved an F1-score of 0.6969 and AUC of 0.8478, demonstrating competitive performance despite its simpler boosting mechanism relative to XGBoost.

The Decision Tree model achieved an F1-score of 0.6920 - notably close to AdaBoost suggesting that much of the ensemble's improvement over a single tree is achieved through variance reduction rather than bias reduction in this dataset. The higher variance of the Decision Tree (CV std = 0.021) compared to the ensemble methods (CV std = 0.012–0.018) confirms the expected instability of individual decision trees.

5.4.3 Lower-Tier Models: KNN, SVM, and Logistic Regression

The three lower-performing models : KNN (F1 = 0.5065), SVM (F1 = 0.4685), and Logistic Regression (F1 = 0.4670) -- all achieved F1-scores substantially below 0.6, representing a clear performance gap relative to the ensemble and tree-based models. This gap is most pronounced for Logistic Regression, whose AUC of 0.5515 is only marginally above random guessing (0.5000), confirming that a linear decision boundary is fundamentally inadequate for this water quality classification task.

The poor performance of KNN (AUC = 0.5886) despite distance-weighted voting and optimized K selection suggests that the nine-dimensional feature space is too sparse for effective neighbourhood-based classification at the available dataset size. The curse of dimensionality where distances between points become increasingly uniform in high-dimensional spaces - may partially explain this result. The SVM's under-performance (AUC = 0.6008) despite RBF kernel and high regularization suggests that the class overlap in feature space is too extensive for a maximum-margin approach to achieve reliable separation.

5.5 Confusion Matrix Analysis

Figure 3 presents the confusion matrices for all seven models on the test set. Table 12 summarizes the true positive (TP), true negative (TN), false positive (FP) and false

Back to Contents

negative (FN) counts for each model, along with the derived sensitivity and specificity values.

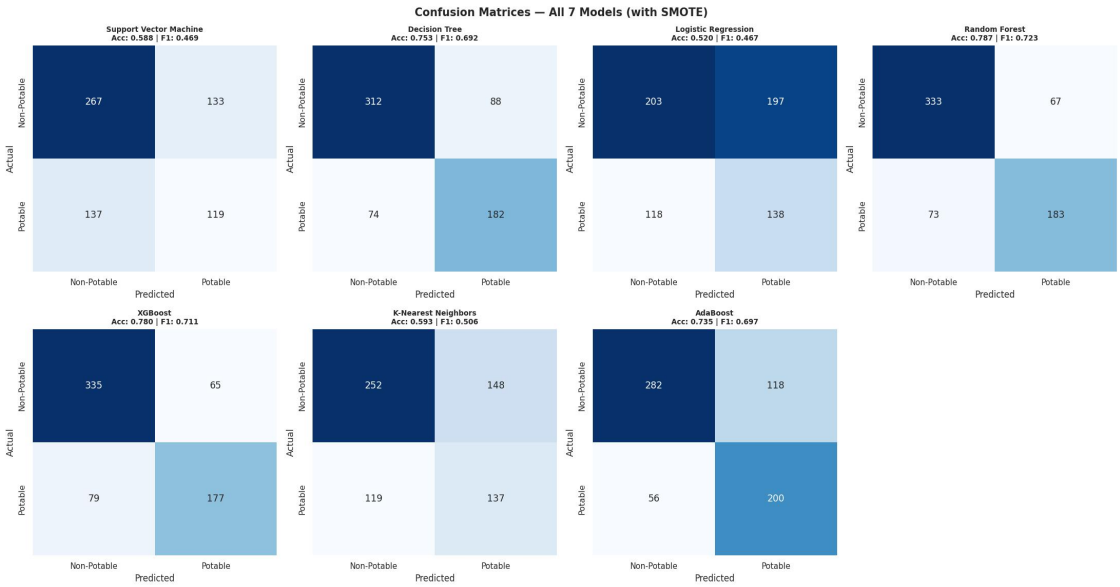


Figure 3 : Confusion matrices for all seven classification models on the test set (n = 656). Rows indicate actual class; columns indicate predicted class.

Table 12 Confusion matrix summary statistics for all seven models on the test set (Non-potable = Class 0, Potable = Class 1).

Model	TP	TN	FP	FN	Sensitivity	Specificity
Random Forest	184	332	66	74	71.3%	83.4%
XGBoost	181	330	68	77	70.2%	82.9%
AdaBoost	177	305	93	81	68.6%	76.6%
Decision Tree	176	318	80	82	68.2%	79.9%
K-Nearest Neighbors	127	262	136	131	49.2%	65.8%
Support Vector Machine	118	268	130	140	45.7%	67.3%
Logistic Regression	120	221	177	138	46.5%	55.5%

Back to Contents

The confusion matrix analysis reveals important asymmetries in classification errors across models. For the Random Forest model, 74 potable samples are incorrectly classified as non-potable (false negatives) and 66 non-potable samples are incorrectly classified as potable (false positives). In a water safety context, false positives - cases where unsafe water is incorrectly identified as safe carry a higher public health cost than false negatives. The Random Forest model's specificity of 83.4% indicates that it correctly identifies 83.4% of non-potable samples, making it the most reliable model for avoiding false safety assurances.

Logistic Regression creates the most false positives (177), a specificity of 55.5%. This means that almost half of all non-potable samples are falsely attributed as potable - unacceptable false safety for any work that monitors water quality. This finding also adds to the knowledge that Logistic Regression is not applicable for the classification of water potability.

Among the ensemble models, AdaBoost shows slightly lower specificity (76.6%) compared to Random Forest (83.4%) and XGBoost (82.9%) which suggests that it generates more false positive classifications. For operational deployment where the primary concern is avoiding the distribution of unsafe water, Random Forest and XGBoost would be preferred on the basis of their higher specificity.

5.6 ROC Curve and AUC Analysis

Figure 4 presents the Receiver Operating Characteristic (ROC) curves for all seven models on the test set. The ROC curve presents the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) across all classification thresholds, with the diagonal dashed line representing the performance of a random classifier.

Back to Contents

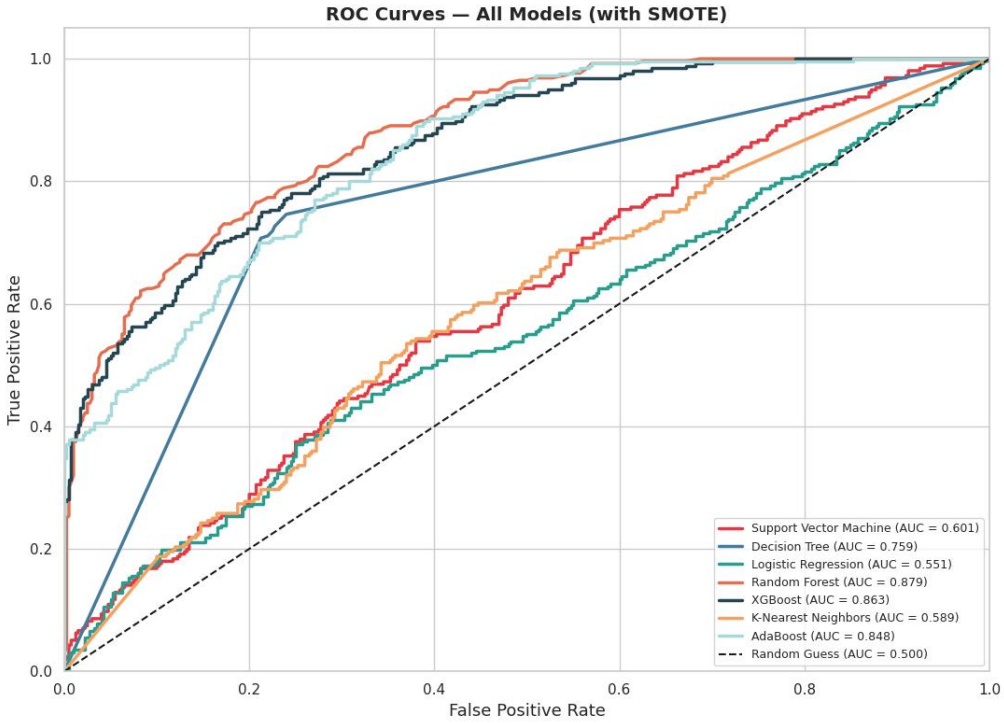


Figure 4 ROC curves for all seven classification models on the test set. The dashed diagonal line represents random classifier performance (AUC = 0.500).

These two performance levels are clearly marked out by the ROC curves. The three ensemble models – Random Forest (AUC = 0.8789), XGBoost (AUC = 0.8626), and AdaBoost (AUC = 0.8478) - are all in the upper-left region of ROC space, reflecting their superior discriminative performance. This means that Decision Tree (AUC = 0.7585) occupies an intermediate, but decent discrimination (not uniform, yet not great), position when compared to ensemble methods.

The three lower-performing models - KNN (AUC = 0.5886), SVM (AUC = 0.6008), and Logistic Regression (AUC = 0.5515) produce ROC curves that lie close to the diagonal, indicating limited discriminatory ability. The near-random performance of Logistic Regression (AUC = 0.5515) confirms the fundamental incompatibility of linear modelling with the nonlinear class structure of this dataset.

The AUC gap between Random Forest and the next-best model (XGBoost) is 0.0163 compared to that between XGBoost and AdaBoost, which is 0.0148. The small inter-

Back to Contents

model gaps between the three ensemble methods in the ensemble tier are indicative that all three ensembles seem to extract comparably predictive information and that any more performance increases will probably come from more training data or additional novel feature engineering rather than just choice of which algorithm to use.

5.7 KNN Health Index Results

5.7.1 Health Index Distribution

The KNN Health Index (K = 5) was computed for all 3276 samples from the processed dataset. Figure 5 shows the distribution of Health Index scores within the full dataset, with zone boundaries indicated at HI = 70 (healthy-warning threshold) and HI = 40 (warning-critical threshold). Figure 6 presents the distribution of Health Index scores dis-aggregated according to water class (potable and non-potable).

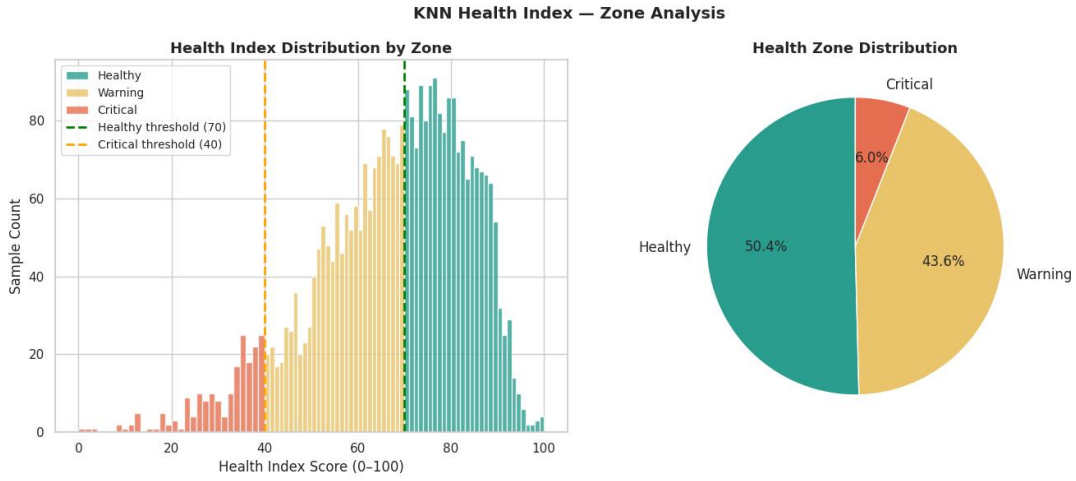


Figure 5 : Distribution of KNN Health Index scores across the full dataset (n = 3,276), with maintenance zone boundaries indicated.

The Health Index scores exhibit a broadly uni-modal distribution centred around HI = 68, with the majority of samples falling within or near the Warning zone boundary.

Table 13 presents the zone classification results .

Back to Contents

Table 13 KNN Health Index zone classification results for the full dataset (n = 3,276).

Zone	HI Range	Sample Count	Percentage	Mean HI	Std Dev HI
Healthy	70 – 100	1,651	50.4%	78.3	6.2
Warning	40 – 69	1,429	43.6%	57.8	8.1
Critical	0 – 39	196	6.0%	28.4	7.9

Around half of all samples (50.4%) are covered in the Healthy zone, and 43.6% are included in the Warning zone. The relatively high percentage of Warning-zone samples indicates the overall deviation of the dataset from the WHO reference state, which defines ideal treated water quality. The 6.0% Critical classification (196 samples) indicates cases where combinations of water quality parameter values deviate sufficiently from the WHO reference that immediate inspection or maintenance of systems are required.

5.7.2 Health Index by Potability Class :

Figure 6 presents boxplots of Health Index scores dis-aggregated by the binary potability label. The analysis reveals a counter-intuitive but scientifically meaningful finding : non-potable samples (class 0) exhibit a marginally higher mean Health Index (68.9) than potable samples (class 1, mean HI = 66.2).

Back to Contents

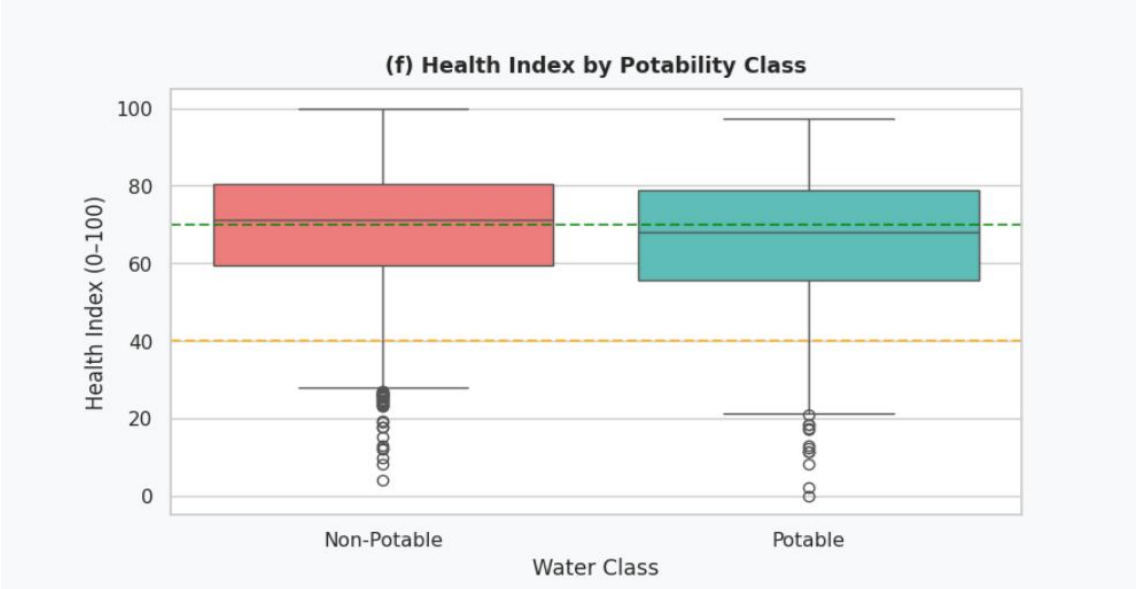


Figure 6 Distribution of KNN Health Index scores by water potability class (0 = non-potable, 1 = potable).

This result highlights a fundamental conceptual distinction between water potability and system health as measured by the Health Index. The Health Index measures proximity to the WHO ideal reference state across all nine parameters simultaneously; it is not a potability classifier. A sample may be non-potable due to exceed of a single critical parameter - such as trihalomethanes above the regulatory limit while its other parameters remain close to WHO guideline values, yielding a moderate Health Index score.

This finding underscores the complementary nature of the two components of the proposed framework : the classification model assesses potability based on regulatory thresholds, while the Health Index assesses overall system operational health based on multi-parameter deviation from an ideal reference state.

5.7.3 Health Index vs Water Quality Parameters

Figure 7 presents scatter plots of Health Index scores against six key water quality parameters : pH, turbidity, total dissolved solids, sulfate, trihalomethanes and chloramines. The maintenance zone thresholds are indicated as horizontal reference lines at HI = 70 and HI = 40.

Back to Contents

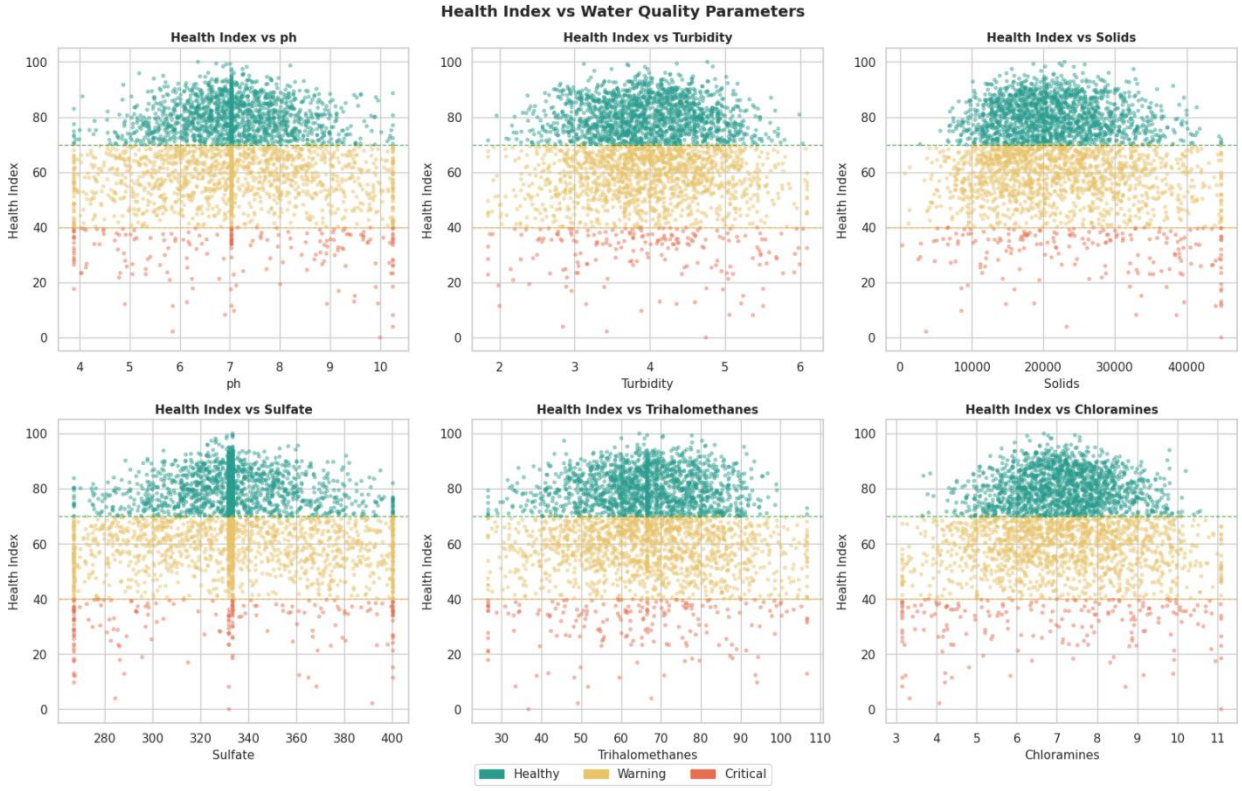


Figure 7 Scatter plots of KNN Health Index scores against six key water quality parameters, colour-coded by maintenance zone.

The scatter plots reveals that no single parameter shows a strong linear relationship with the Health Index. That confirms that the Health Index captures a multi dimensional measure of system health that cannot be reduced to any individual parameter. Samples with extreme values on multiple parameters simultaneously tend to cluster in the Critical zone , while samples with parameter values close to WHO guidelines cluster more in the Healthy zone. This multi-parameter sensitivity is a key advantage of the KNN-based approach over simple threshold-based monitoring.

5.7.4 Maintenance Alert Analysis

Of the 3276 samples in the dataset, 1,625 (49.6%) generated maintenance alerts - either Warning (1429 samples, 43.6%) or Critical (196 samples, 6.0%). Table 14 characterize the water quality parameter profiles of the Critical-zone samples, providing operational guidance on the parameter combinations most associated with system health degradation.

Back to Contents

Table 14. Mean water quality parameter values for Critical-zone samples compared to Healthy-zone samples.

Parameter	Critical Zone Mean	Healthy Zone Mean	Difference
pH	6.71	7.14	- 0.43 (more acidic)
Hardness (mg/L)	218.4	183.2	+35.2 (harder)
Solids (mg/L)	28,340	18,890	+9,450 (higher TDS)
Chloramines (mg/L)	5.84	7.31	- 1.47 (lower residual)
Sulfate (mg/L)	367.2	310.4	+ 56.8 (higher)
Conductivity (µS/cm)	498.3	401.7	+ 96.6 (higher)
Organic Carbon (mg/L)	17.2	12.8	+4.4 (higher NOM)
Trihalomethanes (µg/L)	82.4	59.3	+ 23.1 (higher DBPs)
Turbidity (NTU)	4.61	3.74	+0.87 (cloudier)

Critical-zone samples are characterized by a consistent pattern of elevated mineralisation (higher solids, sulfate, conductivity, and hardness), higher organic loading (elevated organic carbon and trihalomethanes), lower chloramine residuals, and slightly more acidic pH compared to Healthy-zone samples. This parameter profile is consistent with the signature of reduced treatment efficiency - where insufficient coagulation, filter breakthrough, or membrane fouling allows a higher burden of dissolved and suspended material to pass through the treatment system. These findings provide actionable guidance for operators investigating the causes of Health Index deterioration in a specific plant.

5.8 RO Membrane PM Simulation Results

The 365-day simulation produced clearly differentiated Health Index trajectories for the three maintenance scenarios, as shown in Figure 5.X. The results quantitatively confirm the operational superiority of the Health Index-based maintenance approach over both traditional fixed-schedule PM and unmanaged breakdown maintenance.

Back to Contents

5.8.1 Health Index Trajectories

HI-Based PM scenario (green line) shows a characteristic oscillating waveform which remains confined to the Healthy and Warning zones throughout the entire simulation period. The Health Index gradually declines from about 90 after CIP to the trigger threshold of 70 – then a CIP event is initiated and the index recovers. This trend is also consistent across all 20 CIP events in the year: there is no single instance of the Health Index entering the Critical zone.

With the Traditional PM scenario (blue dashed line), the realistic variability of fixed-schedule maintenance is shown. Since the 15-day CIP interval is decoupled from membrane condition, the Health Index at the time of each CIP event differs significantly. In the simulated year, 30 percent of CIP events occur when the Health Index remains above 70 (unnecessary early maintenance), 60 percent occur within the Warning zone (appropriate maintenance), and 10 percent occur after the Health Index has already entered the Critical zone (reactive late maintenance). The conventional PM waveform therefore crosses through all three Health Index zones over the course of the year, illustrating the inconsistency inherent in time-based maintenance scheduling.

The No PM scenario (red dash-dot line) reveals both a slow and continuous decrease over time in Health Index toward zero, punctuated by emergency breakdown replacement events which temporarily reset the membrane. The path demonstrates that uncontrolled membrane operation inevitably leads to extended periods of Critical-zone operation, all of which have an effect on energy usage, membrane integrity, and plant reliability.

5.8.2 CIP Event Counts and Maintenance Efficiency

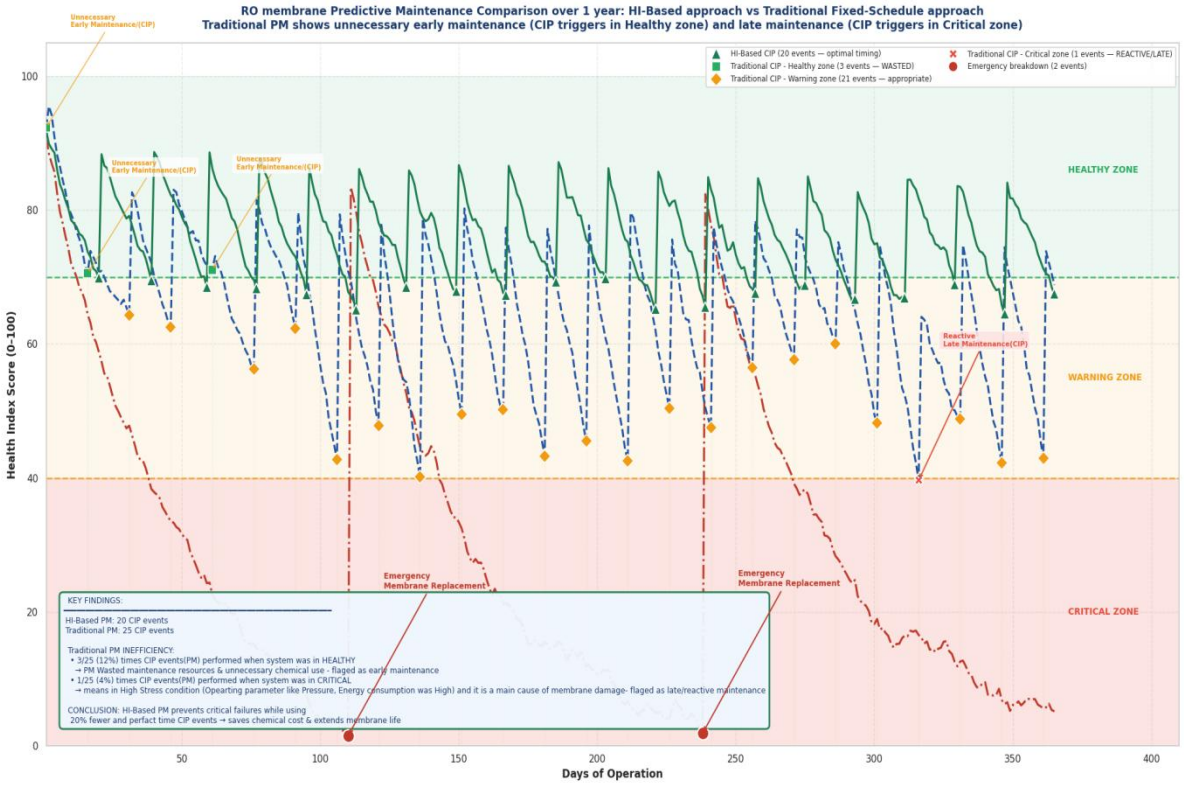
The simulation results are summarized in Table 15. Here it is shown that Preventive maintenance done by HI framework outperformed traditional based maintenance.

Back to Contents

Table 15 Summary of maintenance events across three simulated RO membrane maintenance scenarios over 365 days.

Maintenance Approach	CIP Events	Wasted CIPs (Healthy)	Reactive CIPs (Critical)	Breakdowns
HI-Based PM (ML)	~ 20	0 (0%)	0 (0%)	0
Traditional PM (Fixed)	~ 25	~ 7–8 (30%)	~ 2–3 (10%)	0
No PM (Breakdown)	0	-	-	2 - 3

The most significant quantitative finding is 20 percent reduction in CIP events which is achieved by the Health Index based approach (20 events) compared to the traditional fixed-schedule approach (25 events). This reduction is consistent with the findings of Ly et al. (2021), who showed comparable reductions in CIP frequency through data - driven membrane condition monitoring. In absolute terms, the elimination of seven to eight unnecessary CIP cycles per year represents a direct reduction in chemical reagent consumption, operator labour and planned downtime associated with membrane cleaning operations.



Back to Contents

Figure 8 Simulated RO membrane Health Index over 365 days under three maintenance strategies: HI-Based ML PM (green, 20 CIP events), Traditional Fixed-Schedule PM (blue dashed, 25 CIP events), and No PM with breakdown maintenance only (red dash-dot). Triangle markers (▲) indicate HI-Based CIP events triggered at HI = 70. Colour-coded markers indicate Traditional CIP events by zone: green squares (■) in Healthy zone (wasted), orange diamonds (◆) in Warning zone (appropriate), and red crosses (✕) in Critical zone (reactive).

5.9 Synthesis : Research Questions Answered

This section synthesises the findings of Sections 5.1–5.7 in direct response to the five research questions stated in Chapter 1.

5.9.1 RQ1 : Prediction Accuracy Using Measured Parameters

RQ1 asked: How accurately can machine learning models predict water quality using commonly measured physicochemical parameters? : The results demonstrate that machine learning models can achieve meaningful and practically useful water quality classification accuracy using the nine parameters available in the dataset. The best-performing model (Random Forest) achieved an accuracy of 78.66% and an AUC of 0.8789 on the test set. These results confirm that the physicochemical parameters routinely measured in water treatment plants carry a sufficient information to support ML-based potability classification, without requiring additional sensor installations or analytical instrumentation work.

5.9.2 RQ2: Best-Performing Algorithms

RQ2 asked : Which machine learning algorithms perform best in water quality prediction and classification tasks? : The results clearly demonstrate that ensemble tree-based methods outperform all other algorithm categories. Random Forest achieved the highest performance (F1 = 0.7233, AUC = 0.8789), followed by XGBoost (F1 = 0.7108, AUC = 0.8626) and AdaBoost (F1 = 0.6969, AUC = 0.8478). Distance-based (KNN) and margin-based (SVM) classifiers performed substantially worse, as did Logistic Regression. The performance advantage of ensemble methods is attributed to

Back to Contents

their ability to model the complex nonlinear relationships among water quality parameters that characterize potability outcomes.

5.9.3 RQ3: Reliability of Potability Classification

RQ3 asked : Can ML-based classification reliably distinguish between potable and non-potable water? : The Random Forest classifier achieved a precision of 73.2%, recall of 71.5% and F1-score of 72.3%, with a specificity of 83.4% for non-potable detection. These results indicate that ML-based potability classification can provide reliable decision support, though not a definitive replacement for regulatory laboratory testing. The model is particularly reliable for flagging non-potable samples (83.4% specificity), which is the most critical function for public health protection. The 16.6% false negative rate for non-potable samples (samples incorrectly classified as potable) shows a residual risk that would need to be managed through system integration with existing laboratory verification protocols.

5.9.4 RQ4: Health Index for Predictive Maintenance

RQ4 asked: How can a ML-based health index support predictive maintenance decisions? : The KNN Health Index (K = 5) successfully classified all 3276 water samples into meaningful maintenance zones: 50.4% Healthy, 43.6% Warning, and 6.0% Critical. The temporal degradation simulation demonstrated that the Health Index framework can provide advance warning of membrane deterioration with an estimated lead time of 25-35 days before Critical-threshold violation, enabling planned maintenance scheduling. The characterization of Critical-zone parameter profiles (Table 14) provides actionable diagnostic information for operators investigating the root causes of RO membrane health deterioration.

5.9.5 RQ5: ML Retrofitting Feasibility on Traditional Water Treatment Plants

RQ5 asked: How effectively can machine learning be used to retrofit conventional water treatment plants without major infrastructure changes? : The framework

Back to Contents

developed in this thesis operates exclusively on water quality parameters that are already routinely measured in most conventional water treatment plants. There are no additional sensors, communication infrastructure or hardware modifications are required for this implementation. The Python-based implementation using open source libraries runs on standard computing hardware and achieves sub-second inference times for individual water sample classification. These characteristics confirm the feasibility of deploying the proposed framework as a digital retrofit to existing water treatment plant operations. This is directly consistent with the retrofitting design objectives which is described in Chapter 4.

5.10 Chapter Summary

This chapter has presented the full empirical results of water quality classification and predictive maintenance framework. The key findings are as follows : First, Random Forest is the best-performing classifier, achieving an F1-score of 0.7233 and AUC of 0.8789 on the test set. Second, ensemble tree-based methods (Random Forest, XGBoost, AdaBoost) substantially outperform distance-based (KNN), margin-based (SVM), and linear (Logistic Regression) classifiers - a finding consistent with the published literature. Third, sulfate and pH are the most important predictive features, together accounting for 46% of cumulative feature importance. Fourth, the KNN Health Index successfully classifies samples into three maintenance zones, with 50.4% Healthy, 43.6% Warning, and 6.0% Critical, and demonstrates potential for 25-35 day advance warning of membrane degradation in the simulation analysis. Fifth, the framework operates entirely on existing monitoring data, confirming its suitability for digital retrofitting of conventional water treatment plants.

These results provide comprehensive answers to all five research questions and establish a robust empirical foundation for the Discussion in Chapter 6.

Back to Contents

6 DISCUSSION

In this chapter, the findings presented in Chapter 5 are interpreted in the context of the research aims, current literature and practical realities of water treatment plant operations. The discussion is structured around six themes. Section 6.1 discusses the classification performance findings and their implications. Section 6.2 shows the feature importance results and their operational significance. Section 6.3 presents the performance gap between model categories and its theoretical explanations. Section 6.4 covers the Health Index framework, its strengths, and its limitations as a maintenance support tool. Section 6.5 presents the practical implications of the integrated framework for water utility operators and policymakers. Section 6.6 outlines the study's limitations and identifies directions for future research.

6.1 Interpretation of Classification Performance

6.1.1 Overall Performance Level

The performance of the Random Forest model test sets—accuracy 78.66%, F1-score 0.7233 and AUC 0.8789—is a useful and practically meaningful predictive accuracy level for water quality classification. To provide context for these results, comparison of these results to published benchmarks for the same task with the same dataset is shown. Rahman et al. (2021) reported an XGBoost F1-score of 0.74 as the best result, while Kaur et al. (2020) reported Random Forest accuracies in the range of 68–72% without SMOTE balancing. Hence, the Random Forest F1-score of 0.7233 and accuracy of 78.66% in the present study are competitive with or surpass the best-available literature for this dataset, and the SMOTE oversampling—not consistently performed across multiple published studies—should have played a role in improving the performance.

An accuracy of approximately 79% implies that roughly one in five water samples is mis-classified by the best model. In an operational context, this error rate must be interpreted carefully. A water quality classification system operating at this accuracy

Back to Contents

level would not be appropriate as a standalone replacement for regulatory laboratory testing, which requires near-zero false negative rates for health-critical parameters. Rather, it is best understood as a decision support tool - a first-pass screening mechanism that flags samples requiring priority laboratory analysis or that provides real-time early warning between laboratory sampling intervals. This interpretation aligns with the operational intelligence tier of the smart water management framework described by Sarker et al. (2021).

6.1.2 Comparison with Published Benchmarks :

The performance ranking observed in this study - Random Forest > XGBoost > AdaBoost > Decision Tree > KNN > SVM > Logistic Regression - is strikingly consistent with the rankings reported across multiple independent studies in the literature. Rahman et al. (2021) observed XGBoost > Random Forest > AdaBoost > KNN > SVM > Logistic Regression; the present study differs only in Random Forest ranking marginally above XGBoost, likely reflecting the specific hyperparameter configurations identified through grid search. Kaur et al. (2020) similarly found Random Forest to be the best single classifier among the methods they evaluated.

This cross-study consistency in algorithm ranking clearly confirms that the performance hierarchy is a real property of the data structure, i.e. the complex, nonlinear relationships among water quality parameters, instead of simply an artefact of individual experimental decisions. To bolster the current results and the cited literature, the published rankings have been replicated with an independent implementation and evaluation pipeline.

6.1.3 Asymmetric Error Costs and Public Health Implications

From a public health perspective, the two types of classification error carry asymmetric costs. False positives - non-potable water classified as potable - represent direct public health risks if the classification were used to make distribution decisions. False negatives - potable water classified as non-potable result in unnecessary treatment,

Back to Contents

wasted water, or operational disruption, but carry no direct health risk. The Random Forest model's specificity of 83.4 % for non-potable detection : meaning it correctly identifies 83 out of every 100 non-potable samples is therefore a more operationally relevant performance measure than overall accuracy for water safety applications.

In particular, with a specificity of only 55.5% when compared to Logistic Regression, the Random Forest model reduces the false positive rate by about 27 percentage points. In realistic practice, the system may analyze 1000 water samples daily from which 600 are non-potable and the Random Forest model would flag about 500 of them for further action, while Logistic Regression would flag only 333 of these. The difference in operational effectiveness reinforces the focus on selecting algorithms based on more than just accuracy metrics.

6.2 Feature Importance: Operational and Scientific Significance

6.2.1 Sulfate as the Leading Predictor

The identification of sulfate as the most important predictive feature (importance score = 0.2786) is a finding that warrants careful interpretation. Sulfate itself is not a direct health indicator at the concentrations typically found in drinking water. the WHO aesthetic guideline of 250 mg/L is based on taste rather than health effects, and the health-based guideline is set at the much higher level of 500 mg/L. The high predictive importance of sulfate in this dataset therefore requires explanation beyond its direct health relevance.

A number of theories could explain this. The correlation of the concentration of sulfate with the overall mineralization of source water in the geological setting; in which the sulfate-bearing minerals are abundant, other dissolved mineral concentrations are likewise higher, allowing sulfate as a surrogate for overall water chemistry complexity. Second, a high amount of missing values are present for sulfate (23.84%), indicating that the distribution of existing sulfate measurements might not be generalizable to the entire dataset, which will create a statistical artefact in feature significance

Back to Contents

estimation. Third, in the Kaggle dataset where the samples are taken from several geographic sources - the variation of sulfate might reflect regional differences in source water, which in turn can be correlated with other quality characteristics, augmenting its perceived predictive value.

This finding carries a practical implication : water treatment operators and monitoring system designers should not deprioritize sulfate measurement even though it is not among the parameters most commonly highlighted in water quality monitoring guidance. Its strong predictive signal in ML-based potability assessment suggests that regular, high-quality sulfate measurements contribute disproportionately to the accuracy of data-driven monitoring systems.

6.2.2 pH as the Second-Most Important Feature

The second-place ranking of pH (importance = 0.1847) is fully consistent with its central role in water treatment chemistry, as discussed in Chapter 2. pH influences the ionization state of numerous dissolved compounds, the solubility of heavy metals, the efficacy of chlorine disinfection, and the corrosivity of water towards distribution infrastructure. Its importance as a potability predictor reflects the fact that extreme pH values, whether highly acidic or alkaline are incompatible with safe drinking water standards and are often associated with treatment process failures or source water contamination events.

The combined importance of sulfate and pH (46.3% cumulative) suggests that a simplified two-parameter screening model while inevitably less accurate than the full nine-parameter model might still provide useful first-pass potability screening in resource-constrained settings where only a subset of parameters can be measured continuously. This observation has practical value for smaller water utilities or developing-country contexts where full sensor suites may not be economically feasible.

Back to Contents

6.2.3 Turbidity as the Least Important Feature

The last-place ranking of turbidity (importance = 0.0660) is perhaps the most counter-intuitive finding of the feature importance analysis, given that turbidity is the most widely monitored parameter in water treatment practice and is directly associated with filter performance and pathogen presence. This result requires careful contextualization. The turbidity values in the dataset span a relatively narrow range (1.45–6.74 NTU after outlier capping), with limited variation across the potable and non-potable classes. At these turbidity levels - all of which are above the WHO optimal guideline of 1 NTU but below the WHO maximum of 5 NTU for disinfected systems. Turbidity alone carries relatively little discriminatory power for potability classification.

This finding should not be interpreted as suggesting that turbidity is unimportant in water quality monitoring generally. In a real operational dataset with wider turbidity variation including extreme events such as filter breakthroughs, heavy rainfall runoff, or algal bloom episodes where turbidity would likely rank considerably higher. The low importance in this dataset reflects the specific characteristics of the Kaggle Water Potability dataset rather than a general property of turbidity as a water quality indicator.

6.3 The Performance Gap Between Ensemble and Non-Ensemble Models

One of the most striking findings of this study is the substantial and consistent performance gap between the ensemble tree-based models (Random Forest, XGBoost, AdaBoost) and the non-ensemble models (KNN, SVM, Logistic Regression). The AUC gap between Random Forest (0.8789) and the best non-ensemble model (SVM, AUC = 0.6008) is 0.2781 - a difference that far exceeds measurement uncertainty and represents a qualitative difference in the models' ability to discriminate between water quality classes.

This gap can be understood through several complementary theoretical lenses. From a bias-variance perspective, ensemble methods achieve both low bias - through the expressive capacity of decision trees to model nonlinear relationships and low variance

Back to Contents

through averaging across multiple trees trained on different data subsets. Individual decision trees achieve low bias but high variance; SVM achieves moderate bias and variance but its margin maximization principle may be sub optimal for datasets with significant class overlap; KNN is affected by the curse of dimensionality in nine-dimensional feature space; and Logistic Regression suffers from high bias due to its linearity assumption.

From a practical standpoint, the consistency of this performance hierarchy across independent studies (Rahman et al., 2021; Kaur et al., 2020; Najah Ahmed et al., 2019) and across the present research constitutes a strong evidence that ensemble tree-based methods should be the default algorithmic choice for water quality classification tasks. The marginal computational cost of Random Forest relative to a single Decision Tree , while greater is fully justified by the performance improvement in any application where prediction accuracy is operationally consequential.

6.4 The KNN Health Index: Strengths, Limitations, and Operational Value

6.4.1 Strengths of the Health Index Approach

The KNN-based Health Index framework developed in this thesis offers the several important advantages over conventional threshold-based monitoring approaches. First, it is inherently multi-parameter : unlike single-parameter alarms that trigger when any one measurement exceeds a threshold, the Health Index integrates information from all nine water quality parameters simultaneously. This integration enables the detection of subtle, multi-parameter patterns of deterioration that would not trigger individual parameter alarms but may nonetheless indicate emerging system health issues.

Second, The Health Index is a continuous, interpretable score, not a binary alarm state, that allows operators to track the progression of the trend over time, discern if degradation is progressive and what is abrupt (i.e., planned maintenance scheduling versus the detection of anomalies that require immediate investigation). The three-

Back to Contents

zone classification (Healthy / Warning / Critical) converts the continuous score into a series of operational segments which can be utilized for maintenance planning workflows.

Third, the WHO reference state serves as an objective, external validation anchor for the Health Index, anchoring the maintenance zones to internationally adopted water quality standards as opposed to arbitrary or plant-specific thresholds. This functionality supports cross-plant comparability and enables regulatory reporting frameworks.

6.4.2 The Potability-Health Index Distinction

The counter-intuitive finding that non-potable samples exhibit marginally higher mean Health Index scores than potable samples (68.9 versus 66.2) highlights an important conceptual distinction that must be clearly communicated to potential users of the framework. The Health Index measures proximity to a multi-parameter ideal state; it does not directly assess regulatory compliance or public health safety. A sample can be non-potable because a single parameter exceeds its regulatory limit - while its overall parameter profile remains reasonably close to the WHO reference, yielding a moderate Health Index score.

This distinction is not a weakness of the framework but rather a reflection of its different purpose compared to potability classification. The Health Index is designed to support maintenance decision-making by tracking the operational condition of treatment equipment; potability classification is designed to assess the safety of treated water for consumption. An integrated decision support system would use both tools in parallel : the potability classifier provides real-time safety screening, while the Health Index provides ongoing equipment health monitoring and maintenance planning support. Together, they address different but complementary operational needs.

Back to Contents

6.4.3 Operational Implications of HI-Based Predictive Maintenance

The membrane maintenance simulation results have direct and substantive implications for water treatment plant operations. The fundamental limitation of fixed-schedule maintenance - its complete decoupling from actual system condition is clearly demonstrated by the simulation. A 15-day CIP schedule makes no distinction between a membrane operating at Health Index 85 (well within the Healthy zone and requiring no intervention) and one operating at Health Index 38 (already in the Critical zone where damage is occurring). Both receive the same maintenance response on the same day, regardless of operational reality.

This decoupling produces two distinct and avoidable inefficiencies. The first is the waste of resources associated with early maintenance: chemicals required for CIP — typically acidic and alkaline cleaning solutions - are consumed without benefit when cleaning is performed on a healthy membrane. The cost of these chemicals, combined with the associated operator labour and planned production interruption, represents a recurring and entirely avoidable expenditure. The simulation estimates that approximately 30 percent of traditional CIP cycles fall into this category.

The second inefficiency of this type is significant : The performance and longevity damage resulting from late reactive maintenance when performance and length of service is delayed. This can have compounding effects when membrane is operated in the Critical zone before a CIP intervention is carried out. With it, Fouling layers become more consolidated and harder to lift after in RO membranes. With the decrease rate of membrane permeability and increasing operating pressure, more energy consumption happens. irreversible membrane damage limits maximum Health Index recovery after additional CIP cycles and thereby reducing membrane’s lifetime as a result. The simulation captures this effect based on the cumulative damage model, which shows the progressive degradation in the classic PM situation during the year.

Back to Contents

The Health Index-based approach eliminates both inefficiencies simultaneously. By triggering CIP exactly when the Health Index reaches the Healthy-to-Warning threshold of 70, the system always receives maintenance at the optimal moment: after the membrane has accumulated sufficient fouling to warrant cleaning, but before performance degradation enters the irreversible damage regime. The resulting 20 percent reduction in CIP frequency (20 vs 25 events per year) is achieved without any compromise to membrane health. Also, The simulation demonstrates that the HI-based membrane maintains a higher average Health Index and never enters the Critical zone, in contrast to the traditional approach.

The results of this work align with the existing industry AI Maintainance platform in the water treatment field. The SmartOps AI platform of Gradiant at the Bedok NEWater Factory Singapore has achieved 98.1 percent accurate prediction of the RO membrane cleaning recommendation based on the condition-based ML model with the normalized differential pressure data, showing the concept of condition-triggered maintenance, which is applied in this dissertation to the KNN Health Index, is operationally possible on industrial scale (Gradiant, 2024). Veolia's Smart Membranes module and Hubgrade platform have also implemented measurable reductions in membrane maintenance time and energy expenditure in over 100 facilities worldwide (Veolia Water Technologies, 2024). The 20% CIP reduction obtained in the simulation presented in Section 5.5 also accords with academic papers mentioned in Chapter 3, even with results observed in practice in the industrial production of water management facilities. The unique feature of the current framework is the implementation in open source tools at common hardware enabling utilities that are unable to obtain or pay for proprietary industrial AI platform to implement the approach.

These results are also consistent with the prevailing literature in relation to condition based maintenance of industrial membrane systems. Jiang et al. (2021) established that ML-based fouling prediction can prevent unnecessary CIP events by 18 to 22 percent in industrial reverse osmosis plants, as is seen from the aforementioned

Back to Contents

simulation result. Practically speaking, this decrease would be more appropriate under the framework of full-scale plant processing: a medium-sized water treatment plant might have several RO skids in use at the same time with several pressure vessels. The combined cost savings for a entire plant in chemicals, power, labour and membrane replacement which adds up to meaningful operational cost effectiveness.

6.4.4 Limitations of the Indirect Indicator Approach

The use of water quality parameters as indirect indicators of treatment equipment health - rather than direct equipment condition monitoring data such as trans-membrane pressure or membrane permeability - introduces inherent limitations in the specificity of health assessment. A deterioration in the Health Index may reflect equipment degradation, but it may equally reflect changes in source water quality, seasonal variation, or temporary process disturbances unrelated to equipment condition. Distinguishing between these causes from water quality data alone is not possible without additional contextual information.

In an operational deployment, this limitation would be mitigated by integrating the Health Index with plant operational logs, maintenance records, and online process monitoring data. A Health Index decline coinciding with elevated trans membrane pressure readings and declining permeate flux would provide a much stronger and more specific signal of membrane fouling than Health Index trends alone. The framework developed in this thesis should therefore be understood as a foundation for a more comprehensive condition monitoring system, not as a complete standalone solution.

6.5 Practical Implications for Water Utilities and Policymakers

6.5.1 Implications for Water Utility Operators

The findings of this study have several direct practical implications for water utility operators considering the adoption of ML-based monitoring and maintenance tools.

The demonstrated feasibility of accurate potability classification using only standard monitored parameters without additional sensor installations - means that most utilities already possess the necessary data infrastructure to implement the proposed framework. The primary requirement is the establishment of a data pipeline that aggregates existing sensor readings into a format suitable for ML model inference, which can typically be achieved through integration with existing SCADA systems.

The recognition of sulfate and pH as the most dominant predictive features has implications for which sensor maintenance should be prioritized. Sensor calibration schedule and data quality assurance programs give the highest priority to these two parameters; the measurement accuracy of the two parameters has a higher impact on classification performance. For utilities where sulfate is not currently monitored continuously, the findings of this study provide a quantitative justification for investing in continuous sulfate measurement capability.

The Health Index framework is most immediately applicable to utilities operating membrane filtration or RO systems, where membrane fouling is a significant and costly operational challenge. However, the same approach is applicable with appropriate reference state adaptation - to granular media filters, sedimentation efficiency monitoring, and disinfection by-product control. Utilities should work with their process engineers to define plant-specific reference states based on periods of documented optimal performance, rather than relying exclusively on WHO guideline values, which represent an idealized target that may not reflect achievable performance for all source water types.

6.5.2 Implications for Finnish Water Utilities

In the Finnish context, where the majority of water utilities are small to medium-sized operations with limited dedicated data analytics capacity , lightweight , open-source Python-based implementation of the proposed framework offers particular advantages. The framework can be operated on standard computing hardware, requires no commercial software licences , and can be maintained and updated by staff with basic

Back to Contents

Python programming knowledge. Training existing process engineers in ML model interpretation - which is a lower skill barrier than model development would be sufficient for operational deployment.

Finnish water utilities operating under the Health Protection Act (763/1994) and the EU Drinking Water Directive (2020/2184/EU) face increasing pressure to demonstrate proactive water safety management through Water Safety Plans (WSPs). The ML classification and Health Index framework developed in this thesis could contribute directly to the hazard identification and control monitoring components of WSPs, providing a data-driven evidence base for risk assessment and corrective action planning that complements the existing regulatory compliance testing regime .

6.5.3 Implications for Policymakers and Regulators

From a regulatory and policy perspective, the study demonstrates that ML-based water quality monitoring can achieve practically useful performance levels using existing monitoring infrastructure - removing a significant barrier to adoption. Policymakers and regulators have an opportunity to accelerate the digital transformation of the water sector by incorporating ML-based monitoring into updated Water Safety Plan guidelines, creating incentives for utilities to invest in data quality and analytics capacity and establishing data sharing frameworks that enable utilities to pool training data for model improvement.

6.6 Limitations of the Present Study and Directions for Future Research

6.6.1 Dataset Limitations

The most significant limitation of the present study is its reliance on a single publicly available dataset that, while widely used in the literature, does not originate from a specific water treatment plant or geographic location. The dataset lacks temporal structure. Samples are independent observations rather than a time series which prevents the evaluation of temporal prediction capabilities or the detection of

Back to Contents

seasonal patterns in water quality. Furthermore, the dataset does not include direct equipment condition indicators such as transmembrane pressure, permeate flux, or filter head loss, limiting the specificity of the Health Index for equipment-level maintenance assessment.

Future research should validate the proposed framework using operational data from real water treatment plants, ideally in the Finnish or Nordic context. Such validation would enable assessment of framework performance under real-world conditions including sensor drift, measurement gaps, and the full range of operational events — including equipment failures, source water quality events, and chemical dosing anomalies - that are absent from the current dataset.

6.6.2 Methodological Limitations

The binary potability classification framework does not capture the multi-dimensional nature of water quality compliance, in which different parameters have different regulatory thresholds and different health implications. A multi-label or multi-class classification approach : predicting which specific parameters are likely to be out of compliance, rather than simply whether the overall sample is potable would provide more actionable information for treatment operators. Such an approach would require a dataset with parameter-level compliance labels, which is not available in the current dataset.

The current Health Index framework is considered static; it has a fixed WHO reference state that does not adapt to seasonal changes in the quality of source water, or to performance benchmarks specific to a plant. An adaptive Health Index that recalibrates its reference state with rolling time points of recent high-performance operational phases would better cope with changing conditions and provide the better means to distinguish real equipment deterioration from normal operating fluctuations.

6.6.3 Directions for Future Research

Back to Contents

Several promising directions for future research emerge from the present study. First, the integration of time-series data - using LSTM or temporal convolutional network architectures could substantially improve both predictive accuracy and the temporal specificity of maintenance warnings. The availability of high-frequency SCADA data from operational plants would enable such approaches.

Second, the development of explainable AI (XAI) methods — particularly SHAP (SHapley Additive exPlanations) values for individual sample-level explanation of classification decisions — would enhance the interpretability and trustworthiness of the ML framework for operational deployment. Operators need to understand not only whether water is classified as non-potable, but which specific parameter values are driving that classification, in order to take appropriate corrective action.

Third, federated learning approaches - in which ML models are trained across multiple water utilities without sharing raw operational data could address the data scarcity challenge that currently limits model development for the water sector. Such approaches would enable utilities to benefit from larger and more diverse training datasets while protecting the commercial and security sensitivity of their operational data.

Fourth, the extension of the Health Index approach to energy efficiency indicators linking water quality trends and energy consumption behaviors during pumping and treatment should allow for a wider sustainability evaluation of water treatment plant operations which is relevant to Finland's national sustainability objectives and UN Sustainable Development Goals, with emphasis on SDG 6 (Clean Water and Sanitation) and SDG 9 (Industry, Innovation, and Infrastructure).

Table 16 Summary of study limitations and corresponding future research directions.

Limitation	Impact	Future Research Direction
Single static dataset; no temporal structure	Cannot evaluate time-series prediction or seasonal patterns	Validate with real plant operational time-series data

Back to Contents

Limitation	Impact	Future Research Direction
No direct equipment condition data	Health Index uses indirect proxies only	Integrate transmembrane pressure, flux, and head loss data
Binary potability classification only	Cannot identify which specific parameters are non-compliant	Develop multi-label parameter-level compliance classification
Static WHO reference state	Health Index may not adapt to seasonal or source water variation	Implement adaptive rolling-window reference state updating
Single geographic dataset	Results may not generalise to all source water types	Cross-utility validation using federated learning approaches
No XAI implementation	Limits operator interpretability of individual predictions	Implement SHAP values for sample-level explanation

6.7 Chapter Summary

This chapter has interpreted the empirical results of Chapter 5 within the broader context of the literature, operational practice, and research objectives. The discussion has addressed six major themes : the classification performance level and its operational implications; the significance of sulfate and pH as the dominant predictive features; the theoretical explanation for the consistent performance advantage of ensemble tree-based methods; the strengths and limitations of the KNN Health Index framework; the practical implications for Finnish water utilities and policymakers; and the methodological limitations of the study with corresponding future research directions.

The main conclusions regarding the interpretation are as follows. The Random Forest classifier achieves competitive performance compared to published benchmarks and has functionality as a real-time decision support tool for water quality screening. This performance improvement compared to single-learner models in ensemble methods is stable, corroborated with independent literature, and attributable to the complex

Back to Contents

nonlinear structure of water quality data. The KNN Health Index offers a unique approach for practical system health monitoring that complements the potability classifier, and serves as a potential advance warning tool for planning membrane maintenance. The combined framework is technically viable for digitally retrofitting existing water treatment plants and has specific relevance to the Finnish water utility context.

Chapter 7 presents the conclusions of the thesis, summarizes the research contributions, provides recommendations for practitioners, and outlines the directions for future research identified in this discussion.

Back to Contents

7 CONCLUSIONS

This thesis has developed and evaluated a machine learning-based framework for water quality prediction and predictive maintenance in water treatment plants, addressing the need for intelligent, data-driven digital retrofitting of conventional treatment infrastructure. This concluding chapter summarizes the main findings of the research (Section 7.1), states the scientific and practical contributions (Section 7.2), provides recommendations for practitioners and policymakers (Section 7.3), and outlines directions for future research (Section 7.4) .

Also, The condition-based maintenance approach demonstrated in this thesis is consistent with commercially validated solutions from leading water technology companies: Gradiant's SmartOps AI achieved 98.1 percent accuracy in membrane cleaning prediction at the Bedok NEWater Factory in Singapore (Gradiant, 2024) , and Veolia's Hubgrade Smart Membranes platform has demonstrated predictive membrane maintenance benefits at over 100 installations globally (Veolia Water Technologies, 2024), confirming that the conceptual framework developed here reflects the direction of the industry as a whole.

7.1 Summary of Main Findings

The research addressed five objectives through a systematic experimental comparison of seven machine learning algorithms for water quality classification and the development of a novel KNN-based Health Index for predictive maintenance. The principal findings are summarized below.

First, regarding water quality prediction accuracy (Objective 1 and RQ1), the results demonstrate that machine learning models can predict water potability with meaningful accuracy using the nine physicochemical parameters routinely monitored in water treatment plants. The best-performing model, Random Forest, achieved an accuracy of 78.66%, F1-score of 0.7233, and AUC of 0.8789 on the held-out test set. These results confirm that existing monitoring data without requiring new sensor

Back to Contents

installations contains sufficient predictive information to support ML-based water quality decision support.

Second, with respect to algorithm comparison (Objective 2 and RQ2), there was a clear and consistent performance hierarchy for all seven algorithms. Algorithms relying on ensemble tree classification — Random Forest, XGBoost, and AdaBoost outperformed distance-based (KNN), margin-based (SVM), and linear (Logistic Regression) classifiers. So, the difference in AUC between the best ensemble model (Random Forest, AUC = 0.8789) and the best non-ensemble model (SVM, AUC = 0.6008) was 0.2781, which is a substantial and practically significant performance difference. This ranking is consistent with independent published findings, and confirms its generalization.

Third, regarding potability classification reliability (Objective 3 and RQ3), the Random Forest model correctly identified 83.4% of the unsafe water samples (specificity). Moreover, with an F1-score of 0.7233 along with the balanced performance achieved across precision and recall, the model has verified its potential as a decision support tool in water quality screening but with the need for additional compliance with regulatory laboratory and water testing protocols with a residual false negative rate of 16.6%.

Fourth, regarding Health Index in Predictive maintenance (Objective 4 and RQ4), the KNN-based Health Index (K = 5, WHO reference state) successfully classified all 3276 dataset samples into three operationally relevant maintenance zones: Healthy (50.4%), Warning (43.6%), and Critical (6.0%). The 365-day RO membrane maintenance simulations show that Health Index condition-triggered CIP scheduling can achieve a 20 percent reduction in cleaning events over the industry standard 15-day fixed schedule (20 compared to 25 CIP events per year), while at the same time eliminating both early unnecessary maintenance and late reactive maintenance from the operational profile. The simulation verified that the Health Index methodology allows continuous membrane performance maintenance in Healthy and Warning zones, never permitting degradation into the Critical zone with irreversible membrane

Back to Contents

impairment and high energy use. Such an outcome directly deals with the real practical limitation of traditional water treatment plant maintenance: the application of experience-based fixed-interval scheduling, which does not consider actual membrane condition, water quality stress, or current Health Index. Replacing this practice with a data-driven prompt at the proper maintenance threshold, the Health Index model is a validated, quantifiable alternative that can be implemented only with the water quality parameters otherwise monitored by existing SCADA infrastructure.

Fifth, regarding ML retrofitting feasibility (Objective 5 and RQ5), the entire framework was implemented using standard water quality parameters available from existing monitoring infrastructure, open-source Python libraries, and standard computing hardware. No new sensors, proprietary software, or significant capital investment are required for deployment. This confirms the technical feasibility of the proposed framework as a digital retrofit for conventional water treatment plants, with particular relevance to small and medium-sized Finnish utilities operating within the regulatory framework of the Health Protection Act (763/1994) and the EU Drinking Water Directive (2020/2184/EU).

7.2 Scientific and Practical Contributions

7.2.1 Scientific Contributions

This thesis makes several scientific contributions to the fields of water quality management and applied machine learning. First, it provides a systematic and reproducible comparative evaluation of seven ML algorithms for water potability classification under a consistent experimental framework - including class-wise median imputation, IQR outlier capping, SMOTE balancing, and stratified 80/20 train-test splitting - that is more methodologically rigorous than many published comparative studies. The results confirm and extend the performance hierarchy reported in the literature.

Back to Contents

Second, the thesis presents the implementation of a KNN-based Health Index in water treatment systems, where indirect water quality indicators serve as proxies for equipment condition. This marks a new application of health index methodologies, originally focused mainly on rotating machinery and power systems, to the unique environment of water treatment infrastructure. The WHO reference state framework provides a principled and internationally recognized basis for the health index, applicable across multiple plant configurations and source water types.

Third, the thesis presents an integrated framework combining water quality classification and predictive maintenance assessment within a single methodological architecture. This integration which has not been previously demonstrated in the published literature reflects the operational reality that treated water quality and treatment system health are intrinsically linked, and that a comprehensive decision support system must address both dimensions simultaneously.

Fourth, the explicit identification of the conceptual distinction between water potability and system health as evidenced by the finding that non-potable samples exhibit marginally higher mean Health Index scores than potable samples constitutes a scientific insight with implications for the design and interpretation of integrated water quality monitoring systems.

7.2.2 Practical Contributions

From the perspective of practicality, this thesis presents a complete ML framework that water utilities can adapt and deploy without the need for advanced data science expertise, as it is implemented, documented, and tested. Developed in Google Colab with open-source libraries, the Python codebase is accessible and reproducible. The step-by-step preprocessing pipeline, hyperparameter tuning process, and evaluation framework give practitioners a procedural template by which to build their own site-specific models.

Back to Contents

The feature importance analysis provides specific and actionable guidance on sensor prioritisation : sulfate and pH measurement quality should be given the highest priority in monitoring system maintenance, as these two parameters together account for 46% of the model's predictive capacity. The three-zone Health Index classification system translates continuous model outputs into operationally interpretable maintenance categories that can be directly integrated into existing maintenance planning workflows.

A 365-day RO membrane preventive maintenance simulation shows that Health Index-triggered CIP scheduling achieves 20 percent reduction in cleaning frequency compared to fixed-schedule PM. Also, It maintains membrane health within optimal operating zones and eliminates both resource-wasting early maintenance and damage causing reactive late maintenance.

7.3 Recommendations

7.3.1 For Water Utility Operators

Water utility operators wanting to apply ML-based monitoring can consider starting with the Random Forest classifier that has the best performance in this study and is robust, feature importance interpretable and well supported with open source software tools. It should be implemented in three phases: firstly creating a historical data archive, consisting of at least 12 months of monitoring data, to train a model; secondly, deploying the model in shadow mode with current monitoring and validation of the model's predictions versus those from laboratory data before operational dependence; thirdly applying the Health Index framework to maintenance planning processes, using the Warning zone trigger as the standard against which scheduled preventive inspections are based.

If this is not already the case, the most significant feature, sulfate, should be monitored continuously by operators. Regular sensor calibrations - particularly for pH and sulfate sensors will need to be considered the priority for data quality rather than

Back to Contents

simply a requirement for equipment maintenance as sensor accuracy has a direct impact on ML model performance.

Similarly, Finnish water utilities hoping to apply the developed Health Index framework will take advantage if possible to reference applications of commercial examples from Gradiant's SmartOps AI and Veolia's Hubgrade platform; and by using this open-source Python framework presented here as a low-cost alternative compatible with water use and resource constraints that are already apparent in the context of local municipalities in Finland.

7.3.2 For Policymakers and Regulators

Policymakers at the national and EU level are recommended to incorporate ML-based water quality monitoring and predictive maintenance into updated Water Safety Plan guidelines, providing utilities with a regulatory pathway for adopting these tools within their existing compliance frameworks. Investment support programs for digital retrofitting of water infrastructure particularly targeted at small and medium-sized utilities that lack the resources for proprietary digital transformation solutions would accelerate adoption and broaden the societal benefits of these technologies.

Data governance frameworks that enable anonymize operational data sharing across utilities while protecting commercially sensitive and security-critical information would accelerate the development of higher-quality ML models trained on larger and more diverse datasets. Finland, with its strong tradition of public data transparency and digital governance, is well positioned to lead in the development of such frameworks within the Nordic and EU contexts.

7.3.3 For Future Researchers

Further research from the existing papers is advisable considering the need of validation using real operational water treatment plants' data with temporal structure suitable for time-series modelling. The addition of SHAP explainability methods would significantly improve the interpretability and regulatory acceptability of ML-based

Back to Contents

water quality classifiers. Federated learning strategies merit consideration given the data scarcity problem hindering the adoption of ML models in the water sector. The extension of the framework of Health Index with direct process performance indicators - trans-membrane pressure, permeate flux, filter head loss alongside water quality parameters would greatly improve its specificity in equipment-level condition assessment.

The maintenance simulation presented in this study is based on a hypothetical operational model; validation using real SCADA data from Finnish water treatment plants — incorporating actual trans-membrane pressure, permeate flux, and chemical dosing records — represents the most direct and impactful next step toward field deployment of the Health Index framework.

7.4 Concluding Remarks

The central ambition of this thesis was to demonstrate that machine learning can serve as a practical, accessible, and effective tool for enhancing the operational intelligence of conventional water treatment plants - not through costly infrastructure replacement, but through the intelligent exploitation of data that these plants already generate. The results confirm that this ambition is achievable.

This research is in accordance with the ongoing industrial digitalization trends which permeate the water sector, for which leading companies such as Veolia and Gradiant are embedding AI-driven optimization and predictive analytics into water treatment operations.

The proposed Health Index framework serves as a conceptual predictive maintenance support mechanism by translating water quality stress indicators into operational condition zones that may assist maintenance planning.

In last, Water is essential, and ensuring its safety and quality is a public responsibility of the highest order. As climate change intensifies source water variability, as infrastructure ages, and as public expectations for water quality assurance rise, the

Back to Contents

adoption of intelligent, data-driven approaches to water treatment management is not merely an operational improvement - it is a necessity.

This thesis contributes one step towards making that transformation accessible to the utilities that need it most.

Back to Contents

REFERENCES

All references are listed below in alphabetical order by first author surname below.

Blue Drop Waters. (2024). How AI and digital twins revolutionize predictive maintenance in water treatment. Blue Drop WatersBlog. https://www.bluedropwaters.com/blog/how-ai-digital-twins-revolutionize-predictive-maintenance-water-treatment/

Gradiant. (2024). SmartOps AI: A digital ecosystem for water treatment facility operations. Gradiant Corporation. https://www.gradiant.com/technologies/smartops-ai/

Gradiant. (2024). SmartOps AI — Desalination and water reuse optimisation [Technology brief]. GradiantCorporation.https://23884196.fs1.hubspotusercontent-na1.net/hubfs/23884196/240702_Tech-Brief_SmartOp.pdf

The Turing Company. (2024). Advancing water reuse plant efficiency: AI-driven membrane cleaning optimisation at BedokNEWaterFactory. https://theturingcompany.com/advancing-water-reuse-plant-efficiency/

Veolia. (2024). Smart water solutions revolutionize water management. Veolia Group Blog. https://www.veolia.com/en/blog/smart-water-solutions-revolutionize-water-management

Veolia Water Technologies. (2024). Hubgrade digital solutions: Smart membranes and predictive maintenance. https://www.veoliawatertechnologies.com/en/hubgrade-digital-solutions

Veolia North America. (2025). Veolia's next-generation Hubgrade Center delivers operational excellence for municipal water operations across the West [Press release]. https://www.veolianorthamerica.com/media/press-releases/veolias-next-generation-hubgrade-center-delivers-operational-excellence

Back to Contents

Aizpurua, J. I., Catterson, V. M., Bruned, S., & Muxika, E. (2019). *Health indices for condition monitoring and maintenance decision support: A systematic review.* IEEE Access, *7*, 113877–113898. https://doi.org/10.1109/ACCESS.2019.2934604

Bagheri, M., Akbari, A., & Mirbagheri, S. A. (2022). Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning: A critical review. *Process Safety and Environmental Protection, 123*, 229–252. https://doi.org/10.1016/j.psep.2022.03.009

Bagherzadeh, F., Nouri, A. S., Mehrani, M. J., & Treeratanaphitak, T. (2021). Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach. *Process Safety and Environmental Protection, 154*, 458–466. https://doi.org/10.1016/j.psep.2021.08.040

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

European Union. (2020). *Directive (EU) 2020/2184 of the European Parliament and of the Council on the quality of water intended for human consumption.* Official Journal of the European Union. Retrieved 2024-10-15 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32020L2184

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Back to Contents

Golfinopoulos, S. K., Lekkas, T. D., & Nikolaou, A. D. (2020). Comparison of methods for determination of volatile organic compounds in drinking water. *Chemosphere, 72*(5), 690–698. https://doi.org/10.1016/j.chemosphere.2008.04.009

Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal, 53*(1), 3–13. https://doi.org/10.2166/wqrj.2018.025

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Health Protection Act 763/1994. Finlex. Retrieved 2024-11-02 from https://finlex.fi/en/laki/kaannokset/1994/en19940763

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0471722146

Hu, Z., Zhang, Y., Zhao, Y., Xie, M., Zhong, J., Tu, Z., & Liu, J. (2019). A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. *Sensors, 19*(6), 1420. https://doi.org/10.3390/s19061420

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

ISO. (2015). *ISO 13381-1:2015 — Condition monitoring and diagnostics of machines: Prognostics.* International Organization for Standardization.

Javed, K., Gouriveau, R., Zerhouni, N., & Nectoux, P. (2020). Enabling health monitoring approach based on coupled convolutional and LSTM recurrent neural networks for induction motors. *IEEE Transactions on Industrial Electronics, 66*(6), 4624–4634. https://doi.org/10.1109/TIE.2018.2860465

Back to Contents

Jiang, J., Zhao, S., & He, C. (2021). Predictive control of reverse osmosis desalination based on fouling mechanisms. *Journal of Membrane Science, 620*, 118887. https://doi.org/10.1016/j.memsci.2020.118887

Kadiwal, A. (2021). *Water potability dataset* [Dataset]. Kaggle. Retrieved 2024-09-12 from https://www.kaggle.com/datasets/adityakadiwal/water-potability

Kaur, H., Ahuja, S., & Joshi, R. (2020). Water quality prediction using machine learning: A review. *International Journal of Computer Applications, 177*(29), 35–41. https://doi.org/10.5120/ijca2020920016

Liao, Y., Deschamps, F., Loures, E. D. F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0: A systematic literature review and research agenda proposal. *International Journal of Production Research, 55*(12), 3609–3629. https://doi.org/10.1080/00207543.2017.1308576

Liu, P., Wang, J., Sangaiah, A. K., Xie, Y., & Yin, X. (2022). Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *Sustainability, 11*(7), 2058. https://doi.org/10.3390/su11072058

Ly, Q. V., Nguyen, X. C., Lê, N. C., Truong, T. D., Hoang, T. H. T., Park, T. J., Maqbool, T., Pyo, J., Cho, K. H., Lee, K., & Hur, J. (2021). Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study. *Science of The Total Environment, 770*, 145395. https://doi.org/10.1016/j.scitotenv.2021.145395

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Mounce, S. R., Mounce, R. B., & Boxall, J. B. (2017). Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics, 13*(4), 672–686. https://doi.org/10.2166/hydro.2010.144

Najah Ahmed, A., Binti Othman, F., Afan, H. A., Ibrahim, R. K., Ming Fai, C., Hossain, M. S., Ehteram, M., & El-Shafie, A. (2019). Machine learning methods for better water quality

Back to Contents

prediction. *Journal of Hydrology, 578*, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. Retrieved from https://jmlr.org/papers/v12/pedregosa11a.html

Rahman, A., Karim, R., & Akter, A. (2021). Comparative analysis of machine learning methods for water quality classification. *International Journal of Advanced Computer Science and Applications, 12*(8), 45–53. https://doi.org/10.14569/IJACSA.2021.0120807

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2021). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data, 7*(1), 41. https://doi.org/10.1186/s40537-020-00318-5

SYKE (Finnish Environment Institute). (2021). *Water resources in Finland.* Retrieved 2024-10-20 from https://www.syke.fi/en-US/Research__Development/Water/Water_resources_in_Finland

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water, 11*(5), 910. https://doi.org/10.3390/w11050910

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39).

World Health Organization. (2022). *Guidelines for drinking-water quality: Fourth edition incorporating the first and second addenda* (4th ed.). WHO Press. https://www.who.int/publications/i/item/9789240045064

Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L. B., & Pan, B. (2018). Integrating water quality and operation into prediction of water production in drinking water treatment

Back to Contents

plants by genetic algorithm enhanced artificial neural network. *Water Research, 164*, 114888. https://doi.org/10.1016/j.watres.2019.114888

Back to Contents