

CCT College Dublin

ARC (Academic Research Collection)

ICT

Winter 2023

An assessment of the effectiveness of using data analytics to predict death claim seasonality and protection policy review lapses in a life insurance company

Jennifer Loftus
CCT College Dublin

Follow this and additional works at: <https://arc.cct.ie/ict>



Part of the [Computer Engineering Commons](#)

Recommended Citation

Loftus, Jennifer, "An assessment of the effectiveness of using data analytics to predict death claim seasonality and protection policy review lapses in a life insurance company" (2023). *ICT*. 41.
<https://arc.cct.ie/ict/41>

This Capstone Project is brought to you for free and open access by ARC (Academic Research Collection). It has been accepted for inclusion in ICT by an authorized administrator of ARC (Academic Research Collection). For more information, please contact debora@cct.ie.

Applying Machine Learning to Biological Status (QValues) from Physio-chemical Conditions of Irish Rivers

Raúl Martín Sánchez

A Thesis Submitted in Partial Fulfilment
of the requirements for the
Degree of
Master of Science in Data Analytics



September 2023

Supervisor: Kislay Raj

CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Module Title:	Thesis
Assessment Title:	Applying Machine Learning to Predict River Ecological Status from Physio-chemical Conditions of Irish Rivers
Supervisor Name:	Kislay Raj
Student Full Name:	Raúl Martín Sánchez
Student Number:	sbs22021
Assessment Due Date:	22/09/2023
Date of Submission:	22/09/2023
Code repository	https://github.com/sbs22021/wqi
Dashboard	https://biological-status.streamlit.app

RAÚL MARTÍN SÁNCHEZ

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Acknowledgements

To my incredible wife, Isabel, this work would not have come to fruition without her unwavering support. I am deeply thankful for the countless nights you spent beside me, lending a patient ear, offering guidance, and constantly pushing me to do my best and exceed my expectations. Her embodiment of hard work, consistency, and discipline has been a guiding light. With her by my side, the seemingly impossible became achievable, and the daunting became manageable.

I want to thank the scientists, biologists, and experts who generously shared their wisdom and time on this research. Each of you has imparted an invaluable piece of knowledge, shaping this work in profound ways. I earnestly hope that as you read these pages, you recognise the impact of your guidance and see your feedback reflected throughout.

I extend my deepest gratitude to all the lecturers, particularly my project supervisor, Kislay Raj. I also wish to thank the entire staff of CCT College for their invaluable competence and assistance in bringing this project to fruition.

Contents

Acknowledgements	3
Abstract	8
1. Introduction.....	9
2. Research Motivation	11
2.1. Motivation	11
2.2. Problem definition.....	11
2.3. Objectives	14
2.4. Technical Objectives	14
3. Literature Review	15
3.1. Water Quality	15
3.2. Sampling strategies.....	15
3.3. Models for Predicting Water Quality: An Overview	15
3.4. Ecological Status.....	19
3.5. Conclusions	20
4. Methodology	21
4.1. Primary research: In-depth interviews	21
Purposive Sampling	21
Selection Bias.....	22
Primary Research Strategy	23
Depth Interviews.....	23
Bias in Data and its impact on Results	24
Data use for other purposes	24
Ethic Principals.....	25
4.2. Research Methodology.....	27
Business Understanding	28
Understanding.....	28
Data Preparation	28
Modelling.....	29
Evaluation.....	29
Deployment	29
5. Depth Interviews.....	29
5.1. Key Decision Records	32
5.2. Interviewee 1	34
5.3. Interviewee 2	35

5.4. Interviewee 3	36
6. Architecture	37
7. WQI Dataset.....	38
7.1. Data sources	38
7.2. Water level and flows.....	39
7.3. Pollution Impact Potential maps.	40
7.4. Biological Quality Value	42
7.5. Geographical Parameters:.....	42
7.6. Geology and Typology:.....	43
7.7. Size, System, and Protection:.....	44
7.8. Data Dictionary.....	45
7.9. Data storage	47
Chemical Data.....	47
GIS Data.....	47
8. Correlated Chemical Parameters:	48
8.1. Combined Conductivity @20°C and @25°C $\mu\text{S}/\text{cm}$	49
8.2. Total Hardness (as CaCO_3) mg/l	50
8.3. Alkalinity total As CaCO_3 Mg per litre:.....	51
8.4. BOD 5 Days Total Mg per litre.....	51
8.5. Ammonia total As N Mg per litre.....	51
8.6. Dissolved Oxygen % Saturation	51
8.7. Nitrate (as N) mg/l	52
9. Data Preparation	53
10. Models	55
10.1. Common Model Training Approach.....	55
10.2. Naïve Bayes	58
Results	58
10.3. Linear Support Vector Classifier	58
Results	59
10.4. Random Forest Classifier (Best Model).....	59
Model 1 RF: Random Forest Classifier.....	59
Model 2 RF-PCA: Random Forest Classifier with PCA.....	59
Model 3 RF-ERT: Ensemble Random Forest Classifier by River Type.....	59
Model 4 RF-EGT: Ensemble Random Forest Classifier by River Group.....	60
Results	60
RF-01-v0-corr-pips explained:.....	63

10.5.	Multilayer perceptron	66
	Model 1: MLP-01	67
	Model 2: MLP-02	67
	Results:	69
11.	Conclusions	74
12.	Discussion and Future steps.....	75
13.	References	77
14.	Annexes	81
14.1.	Annex I: Interviews	81
14.2.	Request HPC supercomputer	81
14.3.	Chemical parameters	81
14.4.	Model Results (csv).....	86
Abbreviations		86

Tables

Table 1: Interviews Key Decision Records	32
Table 2: Infrastructure	37
Table 3: Data sources	38
Table 4: Data dictionary	45
Table 5: Naive Bayes Results	58
Table 6: SVM Results.....	59
Table 7: Random Forest Results.....	60

Figures

Figure 1: River Ecological Status 2022	10
Figure 2: Classification Of The Status Of Surface Waters And GW according to WFD.....	11
Figure 3 Mathine Learning Models Cloud Tag (Zhu et al., 2022)	16
Figure 4: Theme Analysis.....	30
Figure 5: Interviewee 1 Theme analysis.....	34
Figure 6: Interviewee 2 Theme analysis.....	35
Figure 7: Interviewee 3 Theme Analysis.....	36
Figure 8: Architecture Diagram.....	38

Figure 9: Iris hydrometric gauges (blue) vs river monitoring station locations (red)	40
Figure 10: PIP layer buffer of 500m	41
Figure 11: Ireland River Types	43
Figure 12: Oversampling process for River Types.....	44
Figure 13: QValueID Correlation With Main Parameters	49
Figure 14: Conductivity RS01B010100 @20°C and @25°C (since 2015) From excel	50
Figure 15: Pearson Correlation of Top 7 Parameters with QValueID	Error! Bookmark not defined.
Figure 16: Missing Values DO Series	54
Figure 17: Model Training Steps	57
Figure 18: Random Forest Confusion Matrix (R2 0.7614)	62
Figure 19: Single Decision Tree Graph Visualization Decisions	63
Figure 20: Decision Tree Graph Viz With Zoom	64
Figure 21: Decision Tree Prediction Path - Poor Classification.....	65
Figure 22: Decision Tree, Single Ammonia Second level Leaf example.	66
Figure 23: MLP-02-v0-corr-pips.keras	68
Figure 24: MLP-01-V0-CORR-PIPS.KERAS.....	71
Figure 25: Tensorflow simulation.....	71
Figure 26: MLP Confusion Matrix (Best Model).....	72
Figure 27: MLP Learning Curve (Best Model)	72

Abstract

Water Quality Data

This thesis evaluates and optimises a variety of predictive models for assessing biological classification status, with an emphasis on water quality monitoring. Grounded in previous pertinent studies, it builds on the findings of (Arrighi and Castelli, 2023) concerning Tuscany's river catchments, highlighting a solid correlation between river ecological status and parameters like summer climate and land use. They achieved an 80% prediction precision using the Random Forest algorithm, particularly adept at identifying "good" ecological conditions, leveraging a dataset devoid of chemical data.

Simultaneously, it draws inspiration from (Donohue et al., 2006), who, through their expansive dataset of 797 monitoring stations across Ireland, unveiled pronounced inverse relationships between a river's ecological health and indicators such as urbanisation intensity. Their model, rooted in logistic regression, predicted the likelihood of a river meeting EU Directive's good status criterion with over 75% accuracy, relying on widely available landcover or chemical monitoring data.

In this project and work exploration, the Random Forest Classifier emerged superior, boasting an impressive 83.21% accuracy in its best configuration, with an R^2 value of 0.7614 accentuating its capability. A meticulous feature selection process revealed the efficacy of seven key chemistry parameters: Alkalinity, Ammonia, BOD5, Conductivity, DO, Nitrate, and Total Hardness. This significantly streamlines sampling needs. When these chemical parameters, largely sensor-monitored, are combined with river characteristics and risk indicators such as PIP layers, allowing biological status indicators ahead of the three-year QValue survey timelines, yielding results in merely months or days for processing the latest chemistry results.

The work hereby presented offers a strong foundation in water quality monitoring and model improvement. The findings and methods discussed can help guide future researchers and industry professionals, promoting better decisions and driving progress in refining predictive models and tackling biological classification issues.

Keywords: *Water Quality Prediction, Random Forest, Multi-Layer Perceptron Neural Network (MLP-NN), real-time monitoring, River Water Quality, Biological Status*

1. Introduction

Water quality is a critical factor for environmental and public health, which depends on water's physical, chemical and biological features that determine its use for various purposes. Water is a scarce and valuable resource that human activities affect and needs a holistic management approach in the format of applicable legislation, monitoring from regulatory agencies and policing from local authorities. It also involves social and political aspects such as equity, access and governance. Therefore, improving water quality is essential for environmental sustainability and human well-being.

Water Quality Index (Horton R. K., 1965) is the process of elaborating a numeric value that classifies water quality in simple terms so stakeholders can use it to report and evaluate the results of the water programs in place, among others indicating how clean or polluted a water body is. It also informs the public and policymakers about water quality issues.

Surface water quality can change due to natural and human causes, such as weather, land use, and pollution, making it essential to have reliable and accurate methods to predict surface water quality. This work and data entry require different methodologies, mainly in the form of time series data.

- **Sensor data** (Jian Sha et al., 2021) monitoring parameters can be measured on-site using sensors or probes for Physical and chemical values.
- **Surveys and Monitoring Sampling:** They utilise sample Kits and mostly require Laboratory Analysis. These cannot be measured on-site or need more sophisticated methods or equipment.

This research delves into the utility of diverse machine learning algorithms to assess the biological status of river water based on physio-chemical attributes and inherent river characteristics. It further explores methodologies for selecting pertinent parameters to improve prediction accuracy.

The water condition in Ireland is evaluated based on standards set by the Water Framework Directive and other EU water-related laws. As per the latest Draft Riber Basin Report (Department of Housing, Local Government and Heritage, 2022) there are 2718 surface water bodies examined for ecological status, which ranges from high to bad. The ecological status is determined by considering a range of biological, physico-chemical, and hydromorphological quality elements. Physico-chemical elements comprise nutrient levels, pH, temperature, and dissolved oxygen etc., while hydromorphological elements look at river flow, depth variations, and the structure of riverbeds and shores.

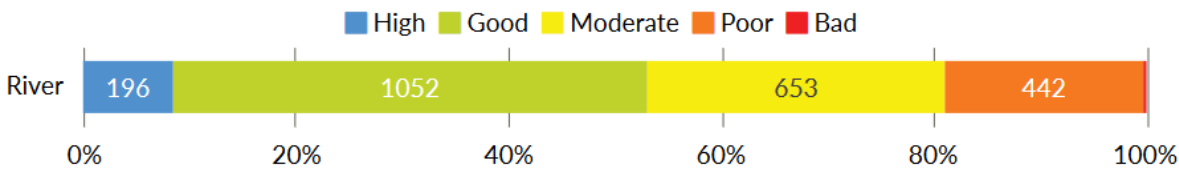


FIGURE 1: RIVER ECOLOGICAL STATUS 2022

(Department of Housing, Local Government and Heritage, 2022)

Biological quality and physico-chemical quality among hydromorphological quality elements determine the Ecological status. A river water body determines this according to the 'one out, all out' principle by the element with the lowest status out of all the assessed biological and supporting quality elements. Overall status assessments consider surface waters' ecological and chemical statuses and their chemical and quantitative statuses.

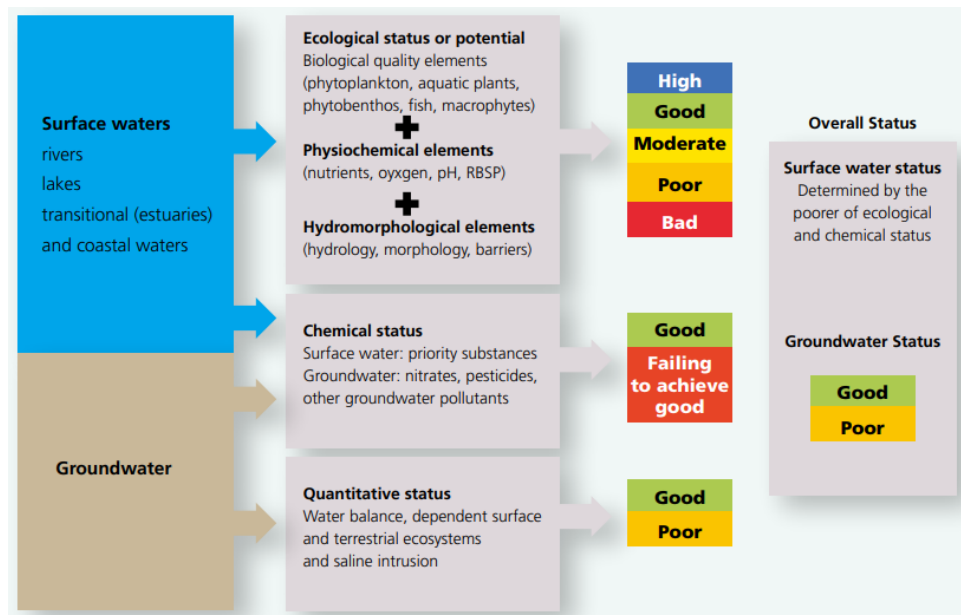


FIGURE 2: CLASSIFICATION OF THE STATUS OF SURFACE WATERS AND GW ACCORDING TO WFD

(O'Boyle et al., 2018, pp. 2013–2018)

2. Research Motivation

2.1. Motivation

River Biological Status in Irish rivers is a critical Water Quality Index surveyed and analysed every three years. Physio-chemical parameters, on the other side, are collected more frequently, and some of those water parameters can be collected in real-time using sensor data.

CAN MACHINE LEARNING MODELS EFFECTIVELY ANTICIPATE RIVERS' BIOLOGICAL STATUS BY INFERRING BIOLOGICAL STATUS FROM PHYSIO-CHEMICAL CONDITIONS, RIVER CHARACTERISATION AND RISK ASSESSMENTS?

2.2. Problem definition

The EU Water Framework Directive aims for all surface waters in the EU to achieve good ecological status, ranging from high to bad (European Parliament, 2000). Various catchment attributes directly influence aquatic systems' water chemistry and ecological status, with

intensified land use, such as urbanisation and agriculture, often leading to decreased ecological quality. In Ireland, the Quality Rating System, primarily based on benthic macroinvertebrate communities, has been employed since 1971 to monitor the ecological quality of over 3000 river sites, proving to be a robust measure linked to chemical status and fish assemblage structures (Donohue et al., 2006).

The River Biological Status in Irish rivers is an essential indicator of water health and quality. The primary metric for this assessment, particularly in Ireland, has historically been analysing macroinvertebrates, including benthic invertebrates (QValues), phytoplankton, fish, macrophytes and aquatic plants (Wilkes et al., 2018).

This choice is rooted in the sensitivity of many macroinvertebrates to environmental disturbances, their relatively localised residence patterns (Struijs et al., 2011), their life cycles that span several months to a year, and the diversity within macroinvertebrate communities. Each species' unique sensitivity to pollutants offers a comprehensive picture of water quality.

However, it is essential to recognise that while macroinvertebrates provide valuable insights, a holistic understanding of aquatic ecosystem health might also consider other biological quality elements, such as fish, macrophytes, and phytoplankton. The broader European guidance, like the Water Framework Directive (WFD), underscores the importance of such a comprehensive approach. Still, practical implementation often requires balancing available resources, historical monitoring practices, and specific characteristics of water bodies.

Physio-chemical parameters provide supplementary information to the biological status. Unlike the triennial nature of biological assessments, these parameters are collected with greater frequency. Advancements in technology have made it possible to monitor some of these parameters in real-time using sensor systems. This frequent and real-time data collection offers a comprehensive overview of current water conditions, allowing for timely and responsive actions.

By analysing the rich dataset of physio-chemical parameters, river characterisations and anthropogenic associated risk, insights into the QValues and potential changes or threats to the biological status of rivers can be anticipated. The rationale behind this is that changes in water chemistry often precede observable shifts in biological communities. In (Zhao et al., 2021), it was found (SO₄²⁻), manganese (Mn), and iron (Fe) concentrations were the water chemistry parameters that best explained bacterioplankton community variation. In the case of invertebrates, this can also be physiological conditions such as temperature (Bonacina et al., 2023).

However, the complex nature of water ecosystems introduces specific challenges to this approach. Water quality is inconsistent; it can fluctuate dramatically based on location, season, and various anthropogenic activities, especially agricultural runoff. The Environmental Protection Agency introduced Pollution Impact Potential (PIP) (EPA Catchments Unit, 2021) maps to recognise this variability and its impacts, which is part of the extensive job produced by the authors (Mockler et al., 2017) regarding phosphorus (P) and nitrogen (N) emissions risks in Ireland. These tool have been instrumental in identifying areas with the highest risk of diffuse phosphorus (P) loss to surface water and diffuse nitrogen (N) loss to both surface and groundwater.

The PIP maps integrate spatial data from farm management, soil types, and hydrogeology by estimating the annual nutrient losses from agricultural landscapes. Instead of being overwhelmed by the vastness of water quality data, these maps provide targeted insights, directing stakeholders towards areas that demand further characterisation and prioritised mitigation efforts.

In the face of these challenges, opportunities for innovation emerge. The diverse nature of water quality data, combined with tools like the PIP maps, paves the way for researchers and scientists to develop advanced forecast models. By harnessing a data-driven approach, there is potential to create an interoperable strategy for water monitoring that considers the unique characteristics of different river systems. Such a strategy could lead to better prediction accuracy, ensuring that water bodies are managed more effectively, and potential issues are addressed proactively.

Yet, these challenges also create room for innovation. The diverse nature of water quality data offers an opportunity for researchers and scientists to develop advanced forecast models. By harnessing a data-driven approach, there is potential to create an interoperable strategy for water monitoring that considers the unique characteristics of different river systems. Such a strategy could lead to better prediction accuracy, ensuring that water bodies are managed more effectively, and potential issues are addressed proactively.

2.3. Objectives

Num	Objective
R01	To innovate the critical water parameters selection and propose methods (standalone and hybrid) to improve predictions effectively, inferring the predicted value of the invertebrate status (QValues).
R02	Apply and propose suitable machine learning models and parameter sets that can accurately predict WQI with different time scales.
R03	Evaluate the potential for cost-effective implementation of the developed model in surface water monitoring practices.

2.4. Technical Objectives

Num	Objective
T01	To compare the performance of different machine learning models in predicting biological status using big data from significant rivers in Ireland.
T02	To determine the specific machine learning models and parameter sets that have the best performance in predicting biological status.
T03	To propose independent variables to be used on the predictive models.
T04	To apply non-linear machine learning techniques to identify the non-linear relationship between features and target variables.
T05	To build a prediction model based on the proposed method and evaluate its accuracy on a water quality dataset.
T06	To compare the accuracy of the proposed method with existing methods with a focus on reducing significant feature selection while maintaining model performance results

3. Literature Review

3.1. Water Quality

The overall water quality status emerges from intricate assessments of numerous indicators, each reflecting different facets of aquatic health and environmental conditions (Figure 2). Such complexity highlights the importance of comprehensive assessments and suggests multiple research avenues to delve into the multifaceted relationships among these indicators.

Over the years, scholars and researchers have embarked on numerous studies focused on distinct aspects of water quality to elucidate the intricate web of interactions shaping our waters' quality status. As the scope of these investigations varies widely, this literature review aims to traverse the breadth of this academic landscape, offering insights from diverse perspectives on water quality assessment.

In doing so, the goal is to eventually hone in on the emerging and promising avenue of utilising machine learning to predict rivers' ecological status (QValues) based on their physio-chemical conditions on Irish rivers. This focus is rooted in the belief that advancements in data analytics and computational methodologies, particularly machine learning, can revolutionise our understanding and prediction capabilities in aquatic ecology.

3.2. Sampling strategies

One of the challenges of data analysis for sensor systems is to balance the trade-off between timeliness and accuracy. The most extended research is typically laboratory samples and mixed real-time monitoring sampling. The former includes more information and parameter details than what can be generated from sensor devices. However, this model provides a more comprehensive and reliable understanding of the System's performance and behaviour while providing timely feedback and alerts for potential issues or anomalies. In addition, sensor data can be applied to parameters such as Dissolved Oxygen (DO) (Wei et al., 2019), Total Nitrogen (TN) (Zhuang et al., 2022), and Total Phosphorus (TP) (Li et al., 2023), which also have the potential benefit of producing acceptable levels of WQI forecasting.

3.3. Models for Predicting Water Quality: An Overview

Different models with distinct assumptions and attributes designed for specific data characteristics present a notable challenge when inter-calibrating and comparing results. Furthermore, this is not just because models have varied characteristics but also because their effectiveness dramatically differs depending on the specific problem and the unique features of the data they handle. It is crucial to understand that no universal benchmark qualifies one model unequivocally superior to another across all scenarios. Indeed, the landscape of modelling is far from being black and white. Instead, it is marked by various shades of grey, where the success of a model often hinges on meticulous preliminary data analyses, iterative testing on target data points, and a deep understanding of external factors influencing the data. The following sections explore various studies that have harnessed diverse models for predicting different water quality indices, shedding light on the intricacies and potential of each, to later on centre the focus on Ecological Status as the focus for this study.



FIGURE 3 MACHINE LEARNING MODELS CLOUD TAG
(ZHU ET AL., 2022)

One significant aspect of this exploration is the classification problem for categorical WQI types.

Tree-based models, encompassing Decision Trees and Ensemble Trees, serve as pivotal tools in this domain. Such models cater to both classification and regression challenges. They offer transparent and easily interpretable pathways for decision-making, demonstrating how various input variables (ranging from pollutant levels to pH, turbidity, and temperature) influence water quality. The ensemble tree approach amalgamates multiple decision trees, aiming to enhance predictive efficacy while curbing the risks of model overfitting.

These methods use primarily two techniques: bagging and boosting. (Sun and Pfahringer, 2011) Bagging creates several datasets from randomly selected training samples, trains decision trees on each, and takes the mean of all predictions. Boosting (Malinin et al., 2021), on the

other hand, adjusts subsequent trees to resolve the net error of the previous tree. (Lu and Ma, 2020) authors propose a hybrid model based on Gradient boosting (XGBoost) and Random Forest (RF) for predicting six water quality parameters, showing that the hybrid model using RF performed the best on temperature, DO, and SC, and Boosting XGBoost showed better performance for pH, Turbidity and FDOM.

As outlined by (Daniels and Koutsougeras, 2021) the KNN approach can be adeptly used for regression tasks, where the values of the 'k' nearest neighbours are either averaged or weighed. For classification, it assigns the new instance the class of most of the k nearest neighbours. This can be useful if similar conditions (temperature, contaminant levels, etc.) produce similar water quality. Authors (A. Danades et al., 2016) used KNN for the water quality status classification, with results indicating that SVM accuracy outperformed the KNN model.

In this regard, researchers also use linear models to decipher the relationships in water quality data. This is the case for Support Vector Machines (SVM), also used for water quality classification tasks (M. Ladjal et al., 2016). In addition, they can help understand the relationship between the parameters and the WQI. These models can be very effective if the relationship is linear or made linear with some kernel trick (as in SVM).

The authors of (Derdour et al., 2022) propose a model based on eleven water quality parameters, obtaining a relevant percentage (95.4%) in water quality classification using SVM algorithms. A similar study of water quality prediction (Islam Khan et al., 2022), consisting of a nine parameters model (pH, DO, COD, TDS, EC, Turbidity, Chloride, SS, and Alkalinity) with PCA and SVR, was found to be the most accurate model for their study, a similar approach used on (Leong et al., 2021) for BOD and COD.

While the primary emphasis of this study focuses on the classification of Ecological Status, it's noteworthy to discuss how time series analysis can be integrated to enhance the prediction.

In (Tan et al., 2012), the authors used a Least Squares Support Vector Machine (LS-SVM) time series prediction model, which leverages phase space reconstruction to transform time series data into vector data, then processed through the LS-SVM model. The LS-SVM-based water quality prediction model outperformed the Backpropagation (BP) and Radial Basis Function (RBF) network methods. Therefore, the LS-SVM method was found to have high predictive accuracy, making it particularly suited for real-time, small-sample water quality forecasts.

On the subject of time series analysis, there is a rising wave of Neural Network models proving to be highly efficient. Such models are commonly employed for sequential data, especially

when dealing with water parameters in a time series format. For instance, (Li et al., 2018) employed Long Short-term Memory (LSTM) to improve the accuracy of predicting Dissolved oxygen in the range from 3 to 12 hours.

A similar study (Yamak et al., 2020) utilised Gated Recurrent Unit (GRU), which can also remember past information to predict future water quality. In particular, researchers have verified that long short-term memory (LSTM) networks and bootstrapped wavelet neural networks (BWNN) can handle fluctuating and nonseasonal time-series water quality (Zhi et al., 2021), (Wang et al., 2013)

In (Lee et al., 2013), a model of 7 parameters is used to study TN and TP, proposing a Multiple Linear regression model for their predictions. (Mohammad Zounemat-Kermani et al., 2019) proposed a model for the DO concentrations prediction problem using Cl, NO_x, TDS, pH, and WT as independent variables. The authors evaluate two approaches: on one side using heuristics models (MLP and CCNN) and a second approach using time-series decomposition, Discrete Wavelet Transform (DWT) and variational mode decomposition (VMD). The analysis found VMD for the combination of NO_x, pH, WT the best algorithm to predict DO.

Having delved into individual models, shifting the focus towards ensemble models is pertinent. By combining different approaches, these models bring a composite perspective that often enhances prediction accuracy and robustness. Ensemble models (Kotu and Deshpande, 2015) combine various models; they can use different algorithms or training datasets and work together to provide a final prediction. The aim is to reduce generalisation error by leveraging the diversity and independence of the base models, effectively capturing the "*wisdom of crowds*". This technique, commonly used in practical machine learning solutions, treats the ensemble of models as a single performing entity. In addition, they can reduce forecasting uncertainty by including predictions from several individual models that use different methods rather than relying on a single one.

In (Shamshirband et al., 2019), ensemble models are employed to make reliable multi-day forecasts of water quality parameters, such as chlorophyll concentration and salinity, in Hilo Bay, Hawaii. The study combined the forecasts of different individual wavelet-artificial neural network (ANN) models using bagging and boosting ensemble techniques. Ensemble models offered the authors better accuracy and reliability in forecasting critical water quality parameters. Individual models were compared with the ensemble models; the latter outperformed the former. For instance, in the case of forecasting chlorophyll levels three days in advance, the ensemble model improved the R² value from 0.75 to 0.81 and reduced the

Root Mean Square Error (RMSE, a measure of the differences between values predicted by a model and the values observed) from 1.80 to 1.36 mg/m³ compared to the best single model. Similar improvements were observed for salinity forecasting.

3.4. Ecological Status

After examining the wide range of models and their roles in predicting water quality, it is essential to narrow the focus to the ecological status of rivers.

The authors of the research on Tuscany's river catchments (Arrighi and Castelli, 2023) unveiled a potent correlation between river ecological status and parameters like summer climate and land use. Notably, the Random Forest algorithm emerged superior, demonstrating an 80% prediction precision for ecological statuses, especially excelling in identifying "good" ecological conditions, utilising a dataset composed of 14 easily accessible features and no chemical data.

In their paper (Donohue et al., 2006) also incorporate into their dataset the interrelations between catchment attributes, the chemical composition of waters, and their consequent impact on the ecological health of rivers.

The dataset consists of 797 distinct monitoring stations spread across Ireland; the authors employ an established biotic index, primarily rooted in the community structures of benthic macroinvertebrates, as their measure of ecological status. The findings from this broad study are striking: there are pronounced inverse relationships between a river's ecological health and several indicators, including the intensity of urbanisation and agriculture within its catchment. The study applies logistic regression, identifying that the most significant pressures compromising the ecological integrity of Irish rivers stem from urbanisation, arable farming practices, and the prevalence of pasturelands within the catchment area. The study reveals that the likelihood of a river meeting the good status criterion laid out by the EU Directive can be predicted with considerably above 75% classification accuracy. This predictive power is harnessed using models employing widely available landcover data or drawing upon chemical monitoring data.

Perhaps the most actionable insights from the paper emerge in the form of non-linear land cover and chemical thresholds. These thresholds can serve as powerful tools for risk management within catchments. The authors' conclusions are both a call to action and a cautionary note: if the current patterns of land use persist, meeting the stringent demands of

the Water Framework Directive will pose a significant challenge. Ensuring the desired water quality will necessitate comprehensive changes, significantly mitigating nutrient exports from agricultural practices. The overarching message is clear: Ireland needs to adopt a more discerning and careful approach to land-use planning to achieve and maintain the water quality benchmarks set by the Directive.

3.5. Conclusions

Over the years, researchers have uncovered the multifaceted interactions of river waters, paving the way for innovative methods like machine learning to predict rivers' ecological status, especially in the context of Irish rivers. This focus aligns with the belief that advancements in data analytics and computational methodologies, particularly machine learning, will significantly influence the aquatic ecology understanding.

Sampling strategies play a pivotal role in balancing timeliness and accuracy. Laboratory samples combined with real-time monitoring sampling offer a holistic understanding of system behaviour, while sensor data proves instrumental in forecasting specific water quality parameters like Dissolved Oxygen (DO), Total Nitrogen (TN), and Total Phosphorus (TP).

Predicting water quality presents many models, each tailored to specific data characteristics. The diverse nature of these models makes it challenging to benchmark their effectiveness universally, as their success largely depends on various external factors and the nature of the data they handle. Both traditional models, like Decision Trees and Ensemble Trees and newer models, such as Support Vector Machines and Neural Networks have proven effective in predicting water quality. Ensemble Models, which amalgamate various models, stand out for their predictive accuracy, especially in forecasting intricate parameters like chlorophyll concentrations.

However, when the spotlight shifts to the ecological status of rivers, it's apparent that specific models exhibit a more significant correlation with environmental conditions. The Random Forest algorithm, for instance, demonstrated significant prediction precision for the ecological statuses of Tuscany's river catchments. The relationship between a river's ecological health and the surrounding environment, particularly urbanisation and agricultural practices, is profound. A transformative approach to land-use planning is paramount to maintain the desired water quality and meet the benchmarks set by directives like the EU Water Framework.

4. Methodology

This section outlines the methodologies employed in this study, focusing on two core areas: primary research through expert interviews and research methodology. The methodologies have been designed to ensure that the data and the study are relevant but also current, compatible, and unbiased.

4.1. Primary research: In-depth interviews

Nonprobability sampling is a method where selection from a population is based on unknown probabilities (Sheppard, 2020); moreover, it does not aim to represent the entire population, but the selections of the samples are not arbitrary, as it is presented next.

Understanding the objectives of this research is indeed central to the selection of an appropriate sampling strategy. This study's primary aim is to predict rivers' biological status using machine learning algorithms based on surface water parameters. This objective involves interpreting and predicting complex, multidimensional data requiring insights from various water perspectives.

The various perspectives are represented by the stakeholder groups involved or populations.

- **Group 1 Legislative:** *Water legislation experts*
- **Group 2 Utilities:** *Water utility professionals*

Each population has been strategically selected (Sheppard, 2020) for its expertise and unique understanding of water quality and its impacts. First, water legislation experts, such as those at the EPA, can provide insights into the regulatory standards and legal framework surrounding water quality. Second, water utility professionals have hands-on experience with water treatment processes and understand the practical challenges of maintaining water quality.

The chosen sampling strategy includes purposive sampling.

Purposive Sampling

The rationale for utilising purposive sampling is driven by the need to gather information from experts with extensive knowledge and experience in water quality and parameters. Reviewing academic literature, professional reports and regulatory documents can identify potential

highly knowledgeable and influential interviewees. This method ensures the information gathered is relevant, detailed, and comprehensive.

Selection Bias

This study's sampling method (Wienclaw, 2021) must provide a representative sample of the population of interest to avoid bias and ensure accurate results. When the sample does not accurately reflect the population, selection bias can lead to incorrect conclusions, such as the infamous 1948 Gallup Poll that wrongly predicted Thomas Dewey's victory over Harry Truman (Lusinchi, 2018). In addition, results from biased samples cannot be reliably extended to the larger population. To minimise the risk of for this, the following conditions will be applied:

- **Random** selection from the pool of scientists in each list will be introduced to remove any possible bias.
- **Diversity:** Include a variety of experts with different water-body domains: rivers, lakes, transitional and coastal.
- **Inclusion Exclusion Critierias:**
 - o At least ten years of experience in the water.
 - o At least 5+ articles and publications relevant to the study.
 - o Group 1: Works or has previously worked in water regulatory agencies.
 - o Group 2: Works or has previously worked in water utilities.

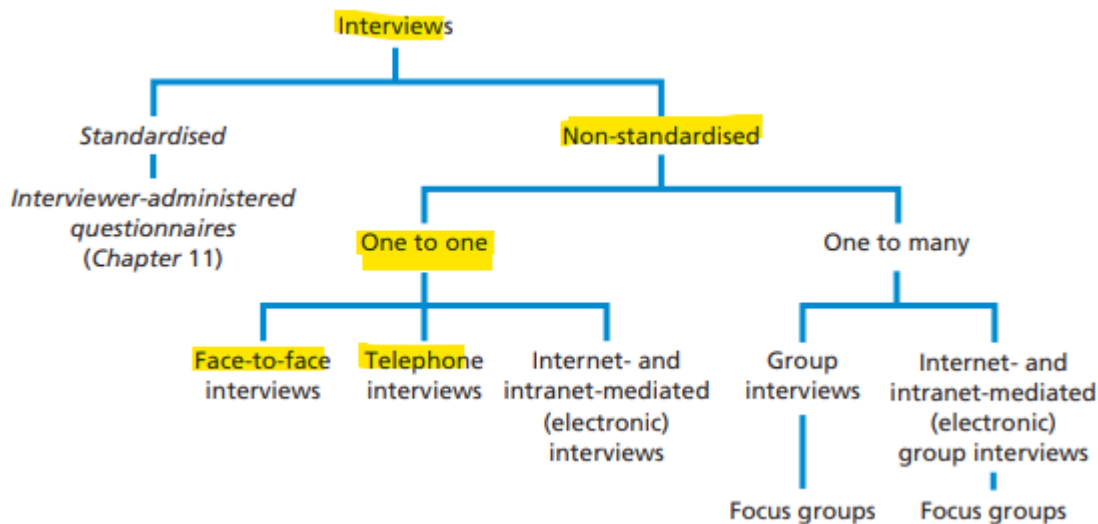
The selection process involves thoroughly reviewing academic literature, professional reports, and regulatory documents to identify individuals who can provide detailed and highly relevant data to inform the research.

A total of 3 in-depth interviews have been organised with experts from various fields related to water quality. The interviews were designed to be open-ended, encouraging participants to share their thoughts, perceptions, and experiences in detail (

Annex I: Interviews).

Primary Research Strategy

The primary research component of this study involves conducting in-depth one-to-one interviews (Saunders et al., n.d.) with selected experts in the two different water domains: legislation and utilities. These interviews aim to gather expert insights on the surface water parameters influencing the WQI and their implications.



Sampling techniques (Saunders et al., n.d.)

Depth Interviews

This strategy aligns with the research objectives and is expected to provide valuable insights into predicting WQI using machine learning algorithms. The following table indicates the allocation of depth Interviews to each group according to the sampling strategy, assuming snowball sampling is required.

Group	Ideal Number of Interviews	Source Options	Purposive Sampling
Water Legislation (EPA)	2	Environmental Protection Agency (EPA) Ireland	1
Water Utilities	1	Water Utilities	1

Each interview of 45 minutes was conducted in a semi-structured (Saunders et al., n.d.) format, allowing for flexibility in the conversation while ensuring key topics were covered and introducing open-ended questions to allow the interviewee to elaborate further details.

The ethical framework for this research is guided by adherence to a defined set of considerations and principles that ensure this study's validity, reliability, and integrity. The principles of informed consent, anonymity, and privacy will be applied from the early stages of data collection. This also extends to the careful and confidential handling and storage of all collected data. Ensuring the research data is free from bias, as it could have substantial undesired consequences and, therefore, must be representative and accurate. The storage and security of the data also demand strict ethical considerations, requiring encryption and restricted access to maintain the confidentiality and integrity of the information.

Upon completing the research, the publication and exploitation of the findings present additional ethical considerations. These include respecting the anonymity of participants in published results, acknowledging all contributors, and being transparent about potential conflicts of interest.

Be mindful of the potential misuse of the study's findings and strive to mitigate this risk. The study aligns with best research practices, fostering credibility and trustworthiness in its results and implications.

Bias in Data and its impact on Results

As indicated in the [Selection Bias](#), bias in data collection can stem from various factors, such as subjective interpretation, non-representative sampling, or biased questioning. When skewed data, the results may not accurately represent the population/group. This ultimately will affect the validity and reliability of the study. E.g. when certain groups of experts are over or under-represented in the sample, their perspectives might not be reflected adequately in the research outcomes. In order to eliminate or minimise this bias in the data collection process, the research must be transparent about the methodology used, adequately representative of the population, and ensure that results are trustworthy and can contribute meaningfully to the study.

Data use for other purposes

While water is a concept that lives in the public domain of our society, the investigation and conclusions from this research may lead to unexpected and uncontrolled secondary uses of the data.

Data must always be preserved and protected to prevent misuse, misinterpretation, or exploitation in ways that could harm the participants or compromise their privacy.

- Obtain informed consent from the participants, clearly stating the purposes of the research and any potential future use of the data.

- Store the data securely using appropriate encryption and access control measures.
By using data encryption and security access protection.
- Anonymise the data, removing any personally identifiable information (PII) that could be used to trace the data back to the participants.
- Share the data with other researchers only when they adhere to ethical guidelines and the original consent conditions.
- GDPR considerations.

Ethic Principals

GDPR

Academic research (Greene et al., 2019) has responsibilities under the GDPR when it requires authors to submit datasets with personal data and subsequently store or process this data, and when published, results could lead to the reidentification of data subjects.

It is also noted that GDPR does not apply when data are anonymous, but this is not true for pseudonymisation. GDPR identifies pseudonymised data in the Regulation as personal data affecting the practices of research studies that consider pseudonymised data non-personal data. In this regard, (Shabani and Borry, 2018) data access control and all other data protection must be applied to these data sets.

Knowing these concepts will help to elaborate on the ethics and considerations for data on this research.

Anonymity and Confidentiality

All published results from primary research should ensure the anonymity and confidentiality of participants. This may involve using pseudonyms or coding systems and removing or altering identifying details. In addition, all participants of in-depth interviews can complete a waiver if they wish to be included in the Acknowledgements section of the final public report.

Exploit Mitigations

The findings of this research will be reported with clarity and Precision. This includes clear definitions of technical terms and avoidance of assumptions wherever possible. Limiting and identifying the context of the study water surface parameters, the primary and secondary data origin (maintaining privacy) and trustable and reliable sources.

Access Control

Access to the encrypted disk with datasets will necessitate a combination of password and biometric authentication (“BIOMETRIC ENABLED ACCESS CONTROL,” 2021) (e.g., fingerprint recognition). This two-factor authentication process ensures an added layer of security for data stored by the researcher and supervisor.

Limited Access

Access to the transactions is restricted to the principal researcher, Raúl Martín Sánchez, the designated supervisor, and professors of the CCT colleague. The principal researcher will be responsible for access control during the research phase. Once the research has been submitted for grading, it will be CCT Colleague and all individuals granted permission to handle this sensitive data and understand their responsibility to maintain its confidentiality.

Audit Trail

Any attempt to access the stored data, whether successful or not, will be logged and monitored. This includes tracking the access time, the identity of the person attempting access, and the files they interacted with. This audit trail contributes to the accountability of the data handlers and serves as a deterrent to unauthorised access.

Data Storage

The transcription and data generated from the depth interviews will be securely stored on an encrypted disk using BitLocker (De Clercq, 2012), a reliable full-disk encryption feature integrated into the Windows operating system that uses Advanced Encryption Standard (AES). It effectively safeguards the data from being accessed or compromised by unauthorised users.

Data Retention and Destruction:

Upon conclusion of the research, the data will be retained for a predetermined period per institutional guidelines or legal requirements (CCT College Dublin, 2020). After this period, all data will be securely and irreversibly destroyed. The method of destruction will adhere to best practices to ensure that no data can be retrieved or reconstructed.

4.2. Research Methodology

In 1996, four leaders of the nascent data mining market (Daimler-Benz, IntegralSolutions Ltd. (ISL), NCR, and OHRA) created CRISP-DM (Costa and Tiago Aparicio, 2020), a methodology focussed on data analysis which involves the realisation of 6 stages in an iterative process. This methodology was applied as part of the research methodology, with the different stages serving as a guiding framework.

The initial stages saw a dynamic preparation process. This entailed a thorough

Literature Review and establishing the current state-of-the-art. Subsequently, in-depth interviews played a crucial role. Incorporating expert feedback with suggestions proved demanding, especially within the project's limited timelines. This integration necessitated further refinement of data, and in some instances, certain propositions, like the exploration of water flow and levels, were sidelined due to the unavailability of historical data.

Business Understanding

During the "**Business Understanding**" phase, the selection of critical water parameters was revised. The goal was to introduce standalone and hybrid methods to enhance predictive accuracy, focusing primarily on inferring the biological status (QValues).

Understanding

The "**Understanding**" phase demanded a rigorous literature review, feedback, and analysis of in-depth reviews to gain a comprehensive grasp of the available data.

The initial two stages were pivotal in enhancing understanding and refining the project's objectives. Such was their impact that some technical objectives underwent reprioritisation, and two research objectives were re-evaluated and replaced.

A notable shift was moving away from the aim of producing predictions on the final chemical pollutants through individual time series analysis. This endeavour warrants its standalone study, leading to the decision to narrow the focus to the classification of the QValue survey exclusively. Furthermore, given that QValues are produced once every three years, providing water stakeholders with estimated QValues results annually or monthly (depending on when chemical results are becoming available) at an accuracy of 83% (see best model) offers immense potential. This can guide improvement initiatives, ensuring that subsequent QValue surveys move in a more informed direction.

Data Preparation

"**Data Preparation**" included dealing with missing data points across stations and addressing null values. Decisions had to be made about datasets that lacked comprehensive historical records across all stations. Data normalisation, column merging, and managing imbalanced river types using oversampling were among the many tasks undertaken.

Consequently, four distinct datasets emerged, capturing aggregated results of prior chemical data since the QValue survey. These datasets spanned different time intervals: 6 months, one year, two years, and an all-encompassing dataset as per Key Decision Records DR3.

By integrating varied combinations, including using PIP layers and focusing on the top 7 correlated chemical parameters or the entire set, 12 unique dataset combinations were curated.

Modelling

The “**Modelling**” phase included a consistent preprocessing of all datasets using a specific pipeline in pyspark. This involved indexing string fields, assembling vectors for categorical values, scaling features (between 0 and 1), and indexing classification statuses (dependent variable). Multiple models were evaluated, including RandomForestClassifier, Forward Multi-layer Perceptron (using pyspark and tensorflow), Naive Bayes Classifier, and Linear Support Vector Classifier. The Feed-Forward Multi-layer Perception, exceptionally executed in TensorFlow (Developers, 2023), exhibited higher versatility due to its parameter adjustments and fine-tuning flexibility.

Evaluation

For “**Evaluation**”, metrics such as accuracy, F1 score, Precision, and recall were generated for model inter-calibration. Visualisation aids like the Confusion Matrix plot and the Learning Curve were also employed in the case of Neural Networks.

Deployment

Finally, in the “**Deployment**” stage, while reserved for the leading researcher, a public dashboard using Streamlit Cloud is available ([Irish River's Biological Status Prediction · Streamlit \(biological-status.streamlit.app\)](https://irishriverbiologicalstatus.streamlit.app)), and upon publishing of this work a facility to allow users to input data to test the predictive capabilities of the finalised models can be discussed.

5. Depth Interviews

A series of in-depth interviews were conducted with experienced biologists. These discussions offered insights into expert and historical viewpoints, including the last 20 years evolving shifts in water parameters and their potential reasons, the direct correlations between chemical changes and observed biological trends and the implications of these shifts on operational processes.

The conversations highlighted the preventative strategies of the EPA, their methodologies for monitoring, and their synergistic endeavours with other entities, as well as exploring the significance of Machine Learning and Artificial Intelligence, evaluating their potential in forecasting the Biological Status and other essential Water Quality Indices.

The transcript has been analysed using thematic analysis (Wæraas, 2022) to identify key themes and subthemes (Figure 4) that are in the scope of the research objectives in order to gain a better understanding and insights related to the surface water parameters and their implications for the WQI, from the perspective of the five domains identified (Monitoring Historic Results, Risks, Water Flow and Pollutants).

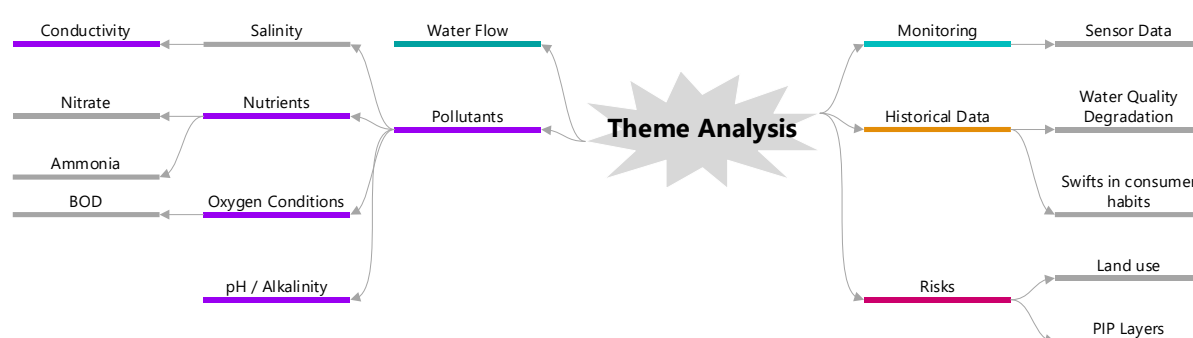


FIGURE 4: THEME ANALYSIS

In the intricate realm of water quality, chemical parameters, rainfall, water flow and other anthropogenic pressures were discussed to unveil how biological status is affected. To address this, discussions took place with three water experts and biologists proficient in the nuances of water quality.

The core motivation behind these interviews lies in uncovering details that could potentially enhance modelling accuracy and highlight aspects of water quality determinants that might traditionally be overlooked. The discussions emphasised the significance and method of integrating chemical data, especially its temporal relevance to the QValue survey, understanding the broader array of water quality determinants like the influence of socio-economic and environmental events, and establishing an exhaustive list of critical pollutants that validate the selection used in this study.

From these sessions, 5 pivotal decisions (*Table 1: Interviews Key Decision Records*) emerged that encompass integrating innovative features such as the Pollution Impact Potential (PIP) Layers, refining the incorporation timeframe for chemical data, recognising historical events influencing water quality trends, and incorporating a comprehensive list of pollutants. The

integration and final model performance are analysed later during the iterative refinement process of the dataset.

5.1. Key Decision Records

TABLE 1: INTERVIEWS KEY DECISION RECORDS

ID	Decision	Source	Rationale	Decision
DR1	Include PIP layers in models.	Interviewee 1	Introduce PIP (Pollution Impact Potential) Layers as features of the training dataset. With current 500 meter buffer definition. This will be expressly used to integrate the risks associated with Nitrate and Phosphorus.	Accepted
DR2	Investigate Water Flow and Hydrometric Gauges	Interviewee 2	Investigate water flow patterns and the current status of hydrometric gauges in Ireland. Given its impact on water flow, the potential inclusion of rainfall analysis will be considered.	Dismissed Due to time constraints and limitations of available public hydrometric gauges and historical data, incorporating this information was not feasible.
DR3	Relevance of Chemical Data	Interviewee 3	Consider different periods to aggregate chemical data with the reference point of the QValue survey. Historical chemical data may not be relevant to QValue survey. Consider the following aggregations. <ul style="list-style-type: none"> - 6 months since QValue survey. - 1 year - 2 years - All available 	Accepted
DR4	Incorporate time context (year)	Interviewee 1,2,3	The model will factor in historical data and significant shifts in trends. This will include events like the 2008 economic downturn, political changes, the removal of EU legislation quotas in 2010, and evolving consumer habits. Including	Accepted

			the year in the model ensures recognition and adjustment for these time patterns.	
DR5	Pollutants	Interviewee 1,2,3	The dataset will assessed for the inclusion of some of the following pollutants: Nutrients (Ammonia, Nitrate, Nitrite), Phosphorus (Total, orthophosphate as P), Oxygen conditions (BOD)	Accepted
DR6	Combine conductivity @20 and @25 degrees	Interviewee 1	There is no significant temperature difference; therefore, it can be considered the same parameter.	Accepted

5.2. Interviewee 1

REGULATORY AGENCY - PHD RESEARCHER LEAD - BIOLOGIST

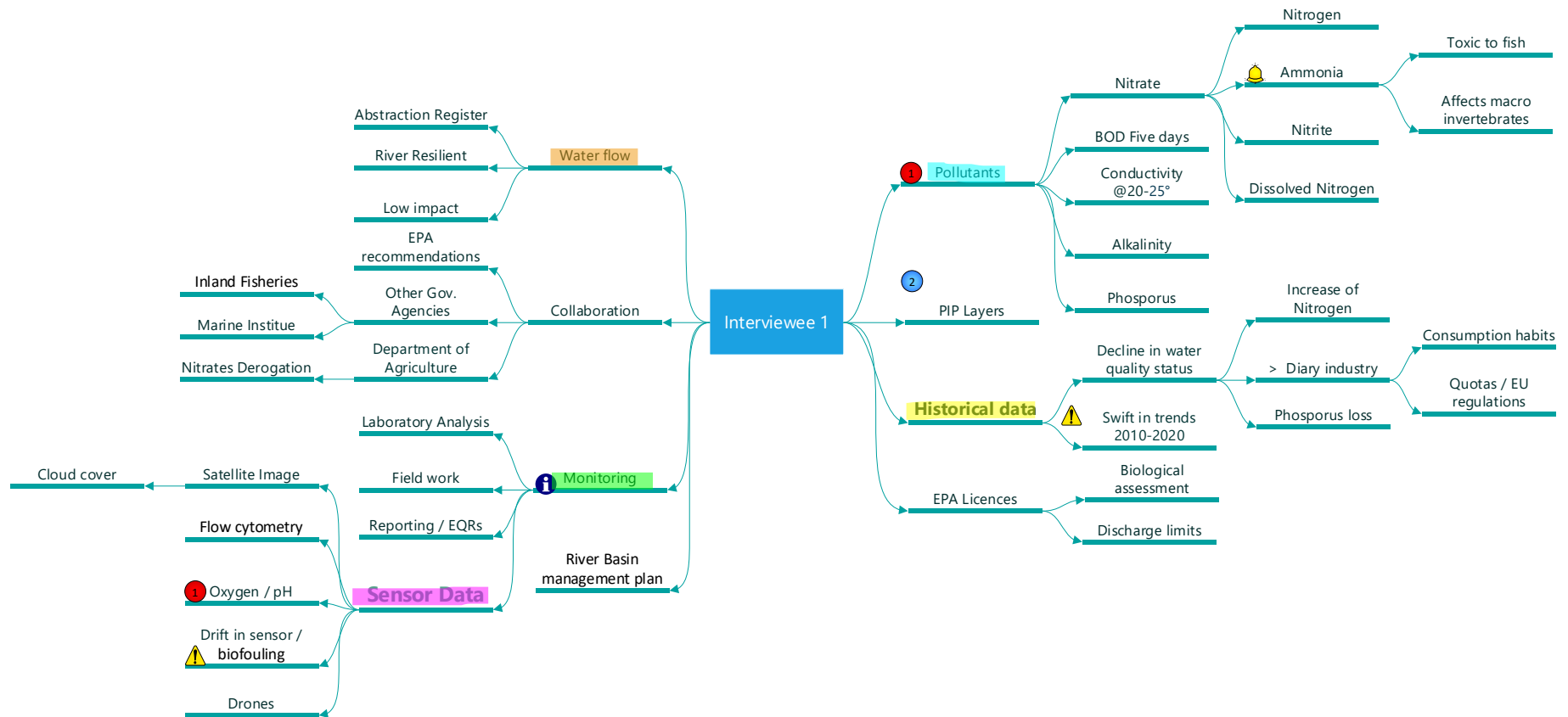


FIGURE 5: INTERVIEWEE 1 THEME ANALYSIS

5.3. Interviewee 2

WATER UTILITIES - PHD STUDENT- RESEARCHER LEAD - BIOLOGIST

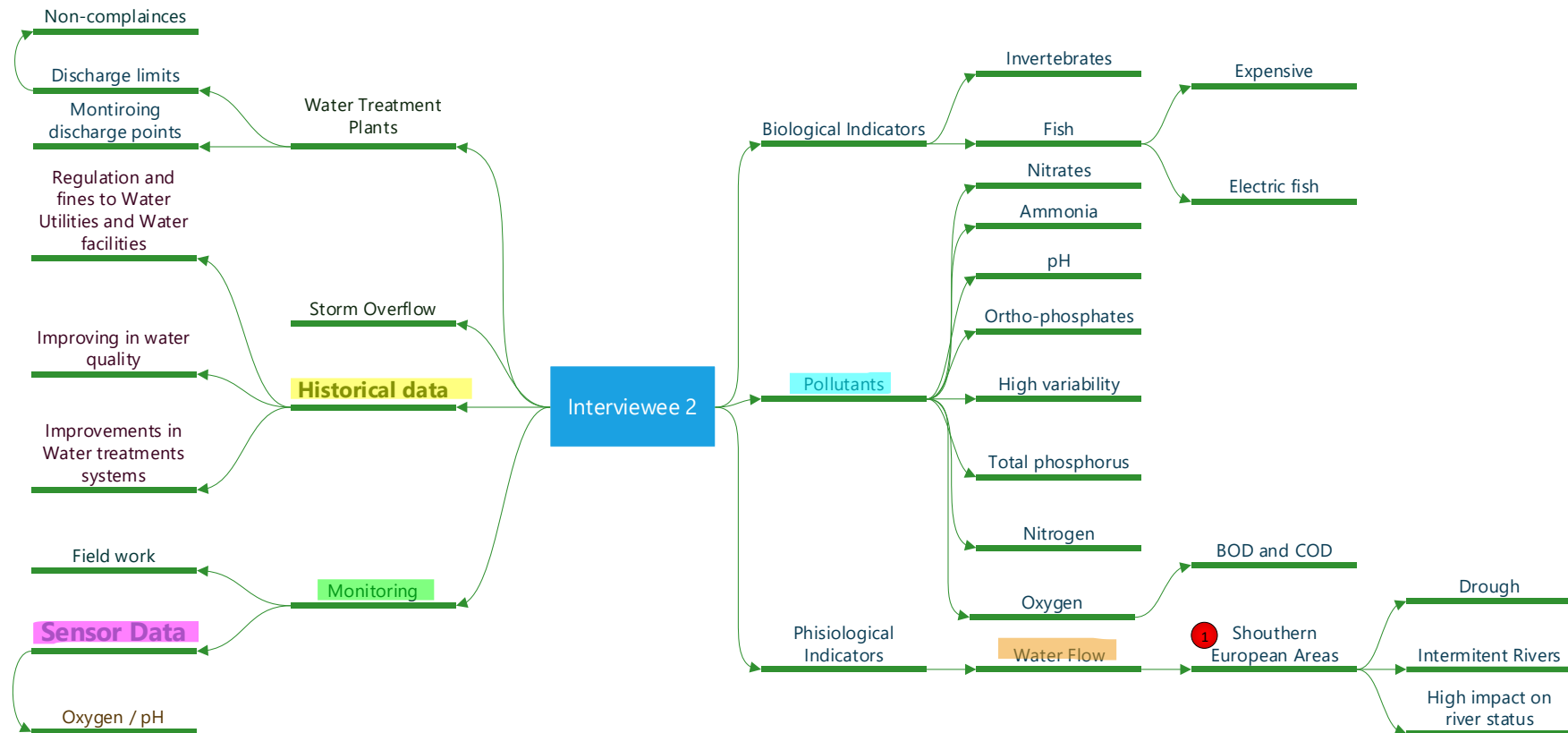


FIGURE 6: INTERVIEWEE 2 THEME ANALYSIS

5.4. Interviewee 3

REGULATORY AGENCY - RESEARCHER LEAD - BIOLOGIST

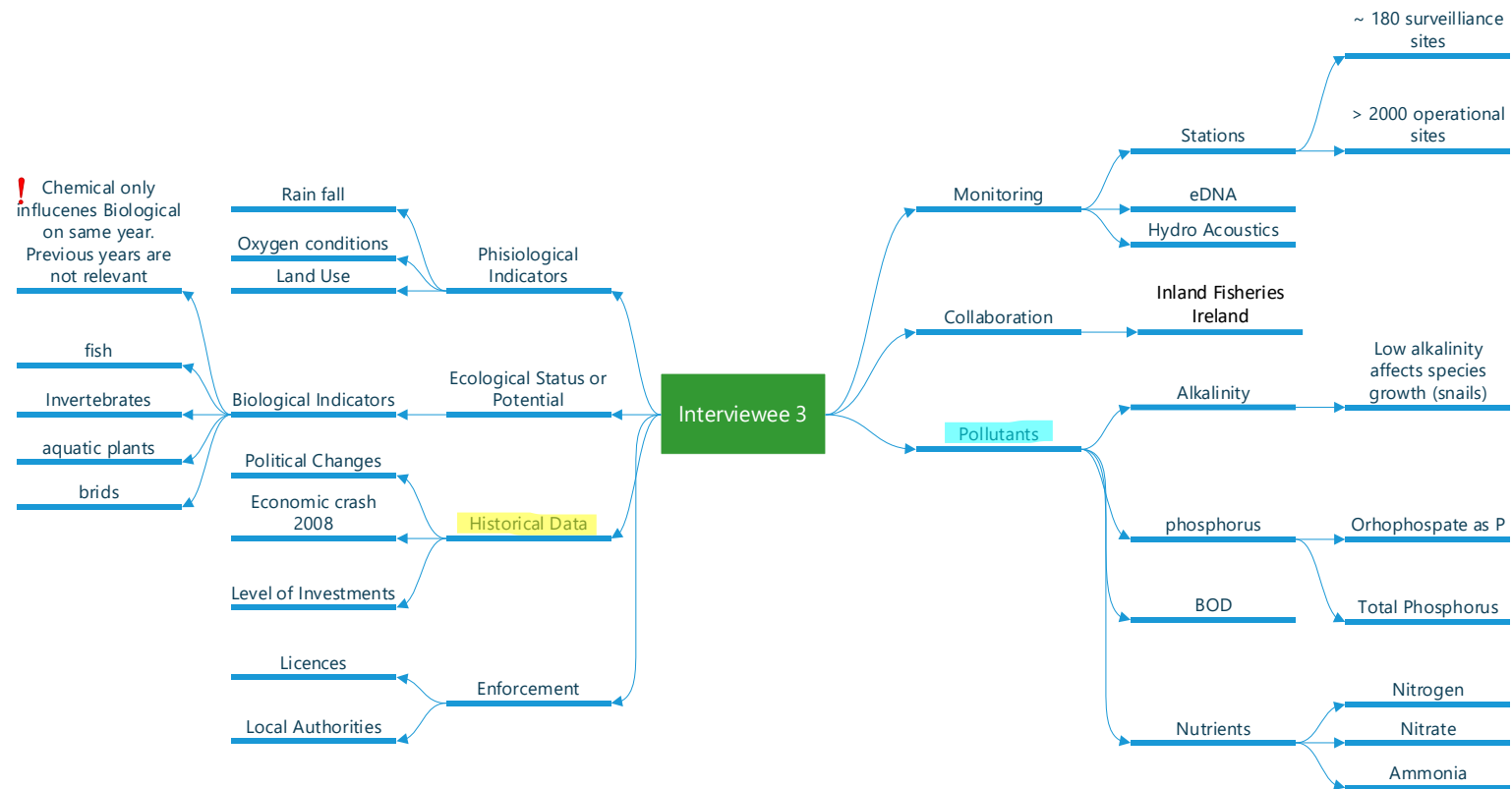


FIGURE 7: INTERVIEWEE 3 THEME ANALYSIS

6. Architecture

This project employed a hybrid architecture of on-premise and cloud hosting, incorporating the following components:

- **Apache Spark VM:** This virtual machine runs Ubuntu 20.04 and comes equipped with Apache Spark 3.2.3 and Jupyter Hub to facilitate big data processing and interactive code execution.
- **Hadoop VM:** This virtual machine operates on Ubuntu 20.04 and features Hadoop Standalone 3.2.4, enabling distributed storage and processing using the MapReduce framework.
- **Research laptop:** Windows 11 professional with Apache Spark and Hadoop.
- **Github:** Code repositor.

TABLE 2: INFRASTRUCTURE

VM Name	Operating System	Software	Version	CPU	Memory	Ports
Apache Spark	VM Ubuntu 20.04 LTS	Apache Spark ("Apache Spark™ - Unified Engine for large-scale data analytics," n.d.)	3.2.3	4	14 GB	8000 (Jupyter Hub)
Hadoop	VM Ubuntu 20.04 LTS	Apache Hadoop (Apache Software Foundation, 2023a)	3.2.4	2	2 GB	9000, 9870, 9864 Hadoop ports
Researcher Laptop	On-premise Windows 11	Apache Spark / Hadoop	Desktop	8	16Gb	n/a
Neo4j Cloud	Cloud	Neo4j Graph DB	Latest	n/a	n/a	443
Streamlit Cloud	Cloud	Streamlit server	Latest	n/a	n/a	443

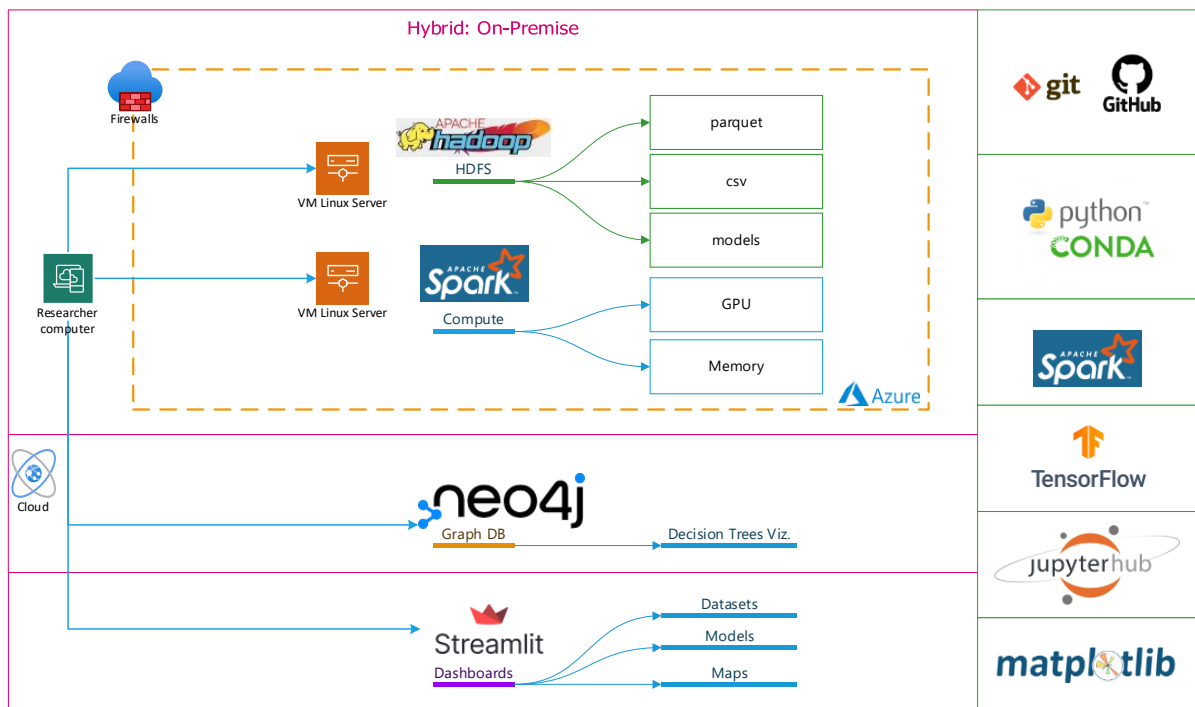


FIGURE 8: ARCHITECTURE DIAGRAM

7.WQI Dataset

The EPA of Ireland provides a rich repository of publicly accessible data for the research community. In this study, multiple resources have been scrutinised. Although the dataset comprises various data categories, particular emphasis has been placed on the Data Flow and PIP layers. While the former layers were not folded into the final dataset, their relevance to the investigation justifies a comprehensive delineation.

7.1. Data sources

TABLE 3: DATA SOURCES

Source	Description	Access to dataset
EPA Geoportal	Monitoring Stations	Mon_QStations_19012023.zip
EPA Geoportal	Hydrometric Gauges	MON_HydrometricGauges_08092022.zip
EPA Geoportal	River Water bodies cycle 2	WFD_RiverWaterbodies_18012021.zip
EPA Geoportal	River Water bodies cycle 3	WFDRiverWaterbodies_27042017.zip
EPA Geoportal	Historical QValues.	Mon_QRecords_19012023.zip

	(dependent variable)	
EPA Catchments	Chemistry Data (EPA Ireland, 2023)	https://wfdapi.edenireland.ie/api
EPA Geoportal	PIP layers	PollutionImpactPotential Data Nov2021.zip
Hydronet	Hydrometric Gauges Monitoring results	https://epawebapp.epa.ie/Hydronet

7.2. Water level and flows.

Data Flow is primarily sourced from real-time hydrometric gauges scattered across Ireland. A holistic list of 2,455 such stations can be accessed at [EPA's GIS portal](#). Real-time and historical data about flow and water levels are catalogued on [Hydronet](#).

This study has discussed this topic with different biologists on how water level and flow rates play an important role in determining the chemical characteristics of river systems. This interaction is especially pronounced in intermittent rivers, which experience alternating wet and dry periods. The flow, or lack thereof, can significantly influence the monitoring results for various reasons:

Concentration Fluctuations:

In intermittent rivers or ephemeral streams, (Gómez et al., 2017) during low flow or dry periods, there might be an increased concentration of certain chemicals due to the reduced volume of water. When flows resume, these concentrations can dilute rapidly, leading to transient spikes or drops in chemical levels.

Sediment Interaction:

The flow rate can impact sediment suspension. Higher flows can resuspend settled particles, which might have absorbed various chemicals. Conversely, during low flows, the settling of particles can remove chemicals from the water column. As discussed with biologists for this study, this is particularly the case for periods of heavy rain or storm water (Nogaro and Mermillod-Blondin, 2009), which is also why they do not take samples during such events, allowing for a few days to settle before a new sample can be taken.

While Ireland might not grapple with the challenges posed by intermittent rivers as acutely as some other European regions, particularly in the south, understanding the dynamics is still

crucial. In regions experiencing more pronounced seasonality or facing increasing water scarcity due to climate change, the intermittency of rivers becomes a significant factor. The chemical oscillations induced by these fluctuations directly impact the ecological status of these water bodies, affecting both the fauna and flora dependent on them.

Therefore, a river's Q-values or ecological status, for instance, certain species, might thrive in higher concentrations of specific nutrients or pH levels. Thus, the interplay between flow, water level, and chemistry determines the river's overall ecological health.

Yet, the intersection of the GIS dataset revealed that only 77 stations offered accessible historical data, whereas the water flow dataset includes 56 stations. Additionally, there is a need for post-processing to align the chemistry monitoring stations with their corresponding hydrometric gauge. The map below illustrates the distance between these points, indicating a necessity for further data refinement. Due to these complexities, these metrics are omitted from the scope of this study. Refer to [Key Decision Records DR2](#).

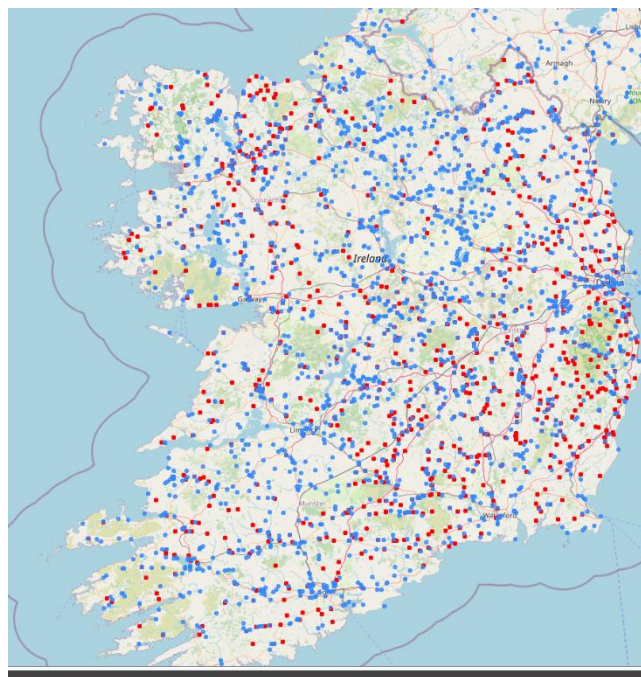


FIGURE 9: IRIS HYDROMETRIC GAUGES (BLUE) VS RIVER MONITORING STATION LOCATIONS (RED)

7.3. Pollution Impact Potential maps.

Features PIP_N, PIP_P, PIP_P_DeliveryPoints, and PIP_P_FlowPaths (EPA Catchments Unit, 2021) are specific risk indicators of nutrient runoff and delivery mechanisms to the water

bodies. This is a direct measure of agricultural impact on water quality, indicating points where nutrient pollution is most likely to enter the water system.

POLLUTANT IMPACT POTENTIAL (PIP) MAPS FOR NITROGEN (N) AND PHOSPHORUS (P) HAVE BEEN GENERATED TO SHOW THE HIGHEST RISK AREAS IN THE LANDSCAPE FOR LOSSES OF N AND P TO WATERS. THESE MAPS, INCLUDING FLOW PATHS AND DELIVERY POINTS, DO NOT INDICATE SPECIFIC AREAS THAT HAVE A PROBLEM, AND THEY ARE NOT DESIGNED OR SUITABLE TO BE USED ON THEIR OWN AS A BASIS FOR DECISIONS AT FIELD SCALE. THEY CAN BE USED HOWEVER TO TARGET MEASURES IN CATCHMENTS WHERE MONITORING DATA HAVE INDICATED THAT THERE IS A PROBLEM

For this study, the pip layers were determined using the following method:

The intersection with the monitoring station and its surroundings is a buffer area of 500 meters, used to identify the highest risk within that zone.

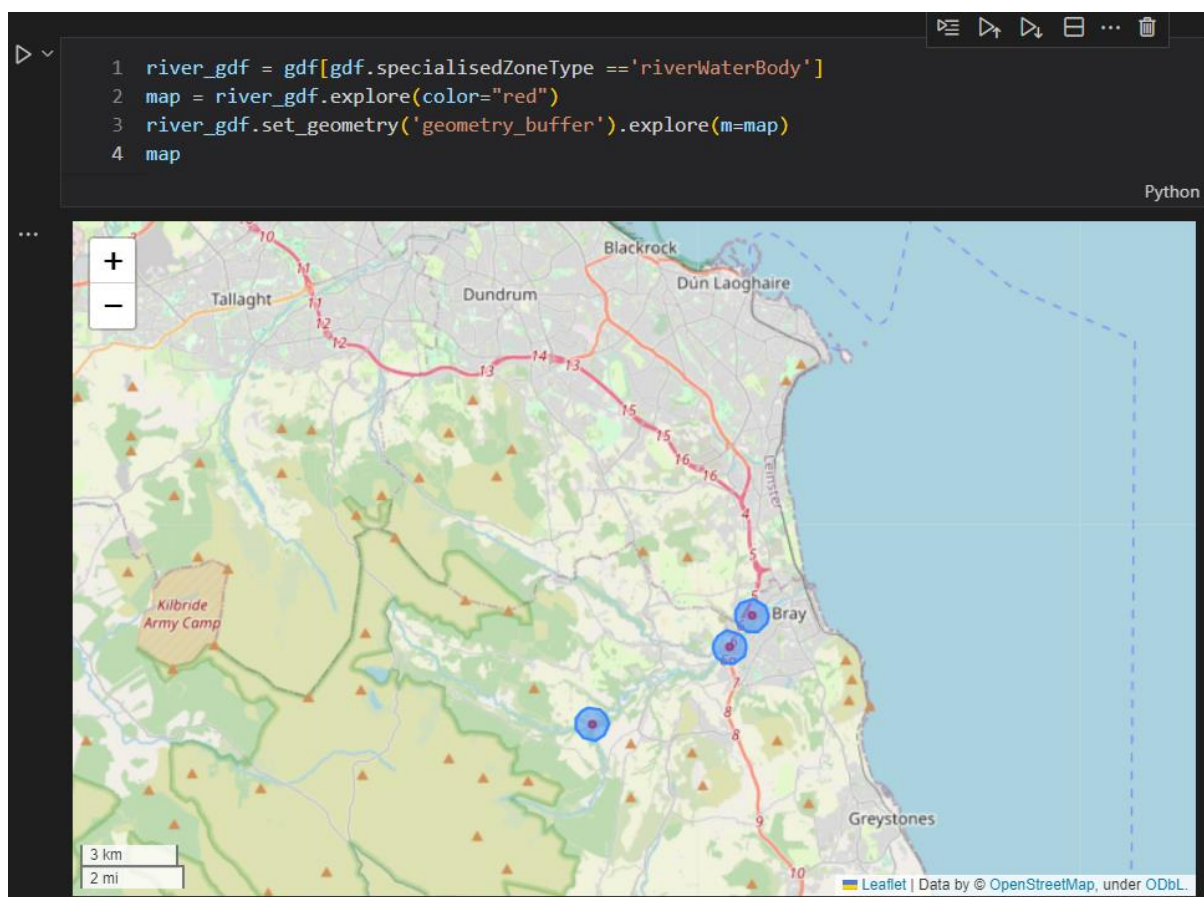


FIGURE 10: PIP LAYER BUFFER OF 500M

This approach, after consultation with biologists, was identified to have certain shortcomings or overlooked several critical aspects:

- River flow direction.
- Potential risks near the monitoring station might not pertain to the station due to their location outside the basin area or because the slope and runoffs do not influence the monitoring station.
- Groundwater infiltration.

A thorough examination of the river's flow direction, main and sub-basin areas, and groundwater infiltration would be essential. However, owing to time constraints and practicality, the PIP layers were produced based on the initial definition. Later model evaluations indicated that including these four features enhanced the accuracy of the models' final dataset by 1 point %, refer to *Model Results (csv)*. While this might not seem like a significant improvement, considering that this data is freely available for relevant periods makes it valuable to include in the analysis.

7.4. Biological Quality Value

The QValueID and QValueScore serve as a direct measure of the biological quality of the river. The former assigns a numerical rating, while the latter provides a descriptive score. This allows for a more nuanced understanding of water quality compared to previously mentioned general classifications.

7.5. Geographical Parameters:

- **Latitude and Longitude** are specific geolocation points that can give spatial context to the water quality data. The authors of (Meyer et al., 2019), indicate that spatial autocorrelated variables may lead to overfitting. When incorporating these properties, outcomes from the test dataset do not seem to validate these concerns. However, another pressing concern is "generalisation." Models can adeptly associate specific outcomes with locations and their histories. Nevertheless, there is no guarantee that monitoring stations will remain consistent in the future, as new stations could be established, existing ones could be replaced, or they might be relocated to different areas. These properties have been excluded from the study to

avoid compromising the model's broader applicability, even at the potential cost of some prediction accuracy.

- **Altitude and slope** might influence water flow and its quality. These geolocational parameters might not have been directly emphasised in the previous papers, but are crucial in understanding spatial variations in water quality.

7.6. Geology and Typology:

- **Geology** and **River Type**: The dataset provides parameters like Geology (Calcareous, Mixed or Silicious) and River Types. These parameters are updated depending on the year of the samples and the associated WFD characterisation cycle. For the current data set, this is WFD Cycle 2 and 3.

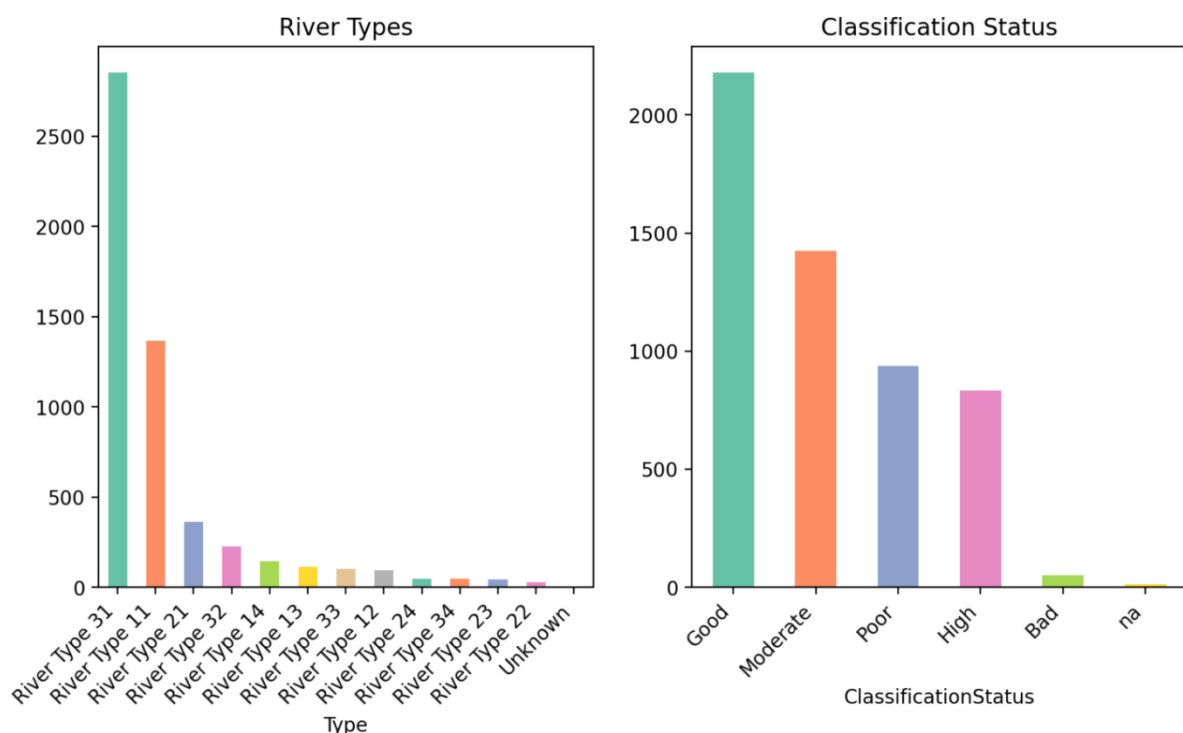


FIGURE 11: IRELAND RIVER TYPES

(KELLY-QUINN ET AL., 2005)

Among all 2,319 monitoring stations in the study, River type 31 emerges as the predominant type, representing 46% of all stations. Following this, River type 11 constitutes 23%, and the

remaining 11 comprise less than 30%. This creates an imbalanced dataset on an important feature that could lead to misrepresentation of the minority of the river types in the models.

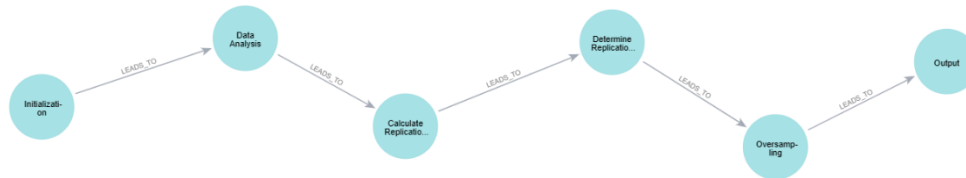


FIGURE 12: OVERSAMPLING PROCESS FOR RIVER TYPES

This oversampling (Mohammed et al., 2020) process is designed to balance out the representation of different river types in the dataset. It first identifies river types not represented as frequently as others and increases the number of entries of these under-represented river types to match those with a higher presence. In essence, it ensures that no river type is under-represented by artificially increasing its count, resulting in a dataset where all river types have a more even representation.

7.7. Size, System, and Protection:

SizeCat and **System** offer insights into the water body's scale and type, which can correlate with water quality. The mention of a **ProtectedArea** can signify areas of ecological or environmental importance.

7.8. Data Dictionary

TABLE 4: DATA DICTIONARY

	Parameter	Description	Type	CAS/EEA Number	QvalueID Correlation
Target	QValueID	Biological Quality Value Identifier: Values corresponding to classification status	Numerical		
	ClassificationStatus	Bad, Poor, Moderate, Good and High	Qualitative / Categorical		
Site characterisation	Year	Year Of QValue Survey occurs	Qualitative / Discrete		
	Geology	Geology: Calcareous, Mixed or Silicious	Qualitative / Categorical		
	Type	River Typology	Qualitative / Categorical		
	SizeCat	Size Category in km2 of the immediate river water body polygon	Numerical		
	System	System A, B or Unknown	Qualitative / Categorical		
	ProtectedArea	Protected Area	Qualitative / Categorical		
	Slope	Slope in metres/metre	Numerical		
	Altitude	Altitude in metres of the water-body's defining point	Numerical		
Chemistry data	AlkalinitytotalAsCaco3Mgl	Alkalinity total As Caco3 (mg/l)	Numerical		-0.356864
	AmmoniatotalAsNMgl	Ammonia total As N (mg/l)	Numerical	CAS_7664-41-7	-0.294154
	Bod5DaysTotalMgl	Biological Oxygen Demand 5 Days Total (mg/l)	Numerical	EEA_3133-01-5	-0.315486
	ConductivityAtXXScm	Conductivity At 20 or 25 Scm	Numerical		-0.400701
	DissolvedOxygenPctSaturation	Dissolved Oxygen % Saturation	Numerical	EEA_3131-01-9	0.310327
	NitrateAsNMgl	Nitrate (as N) mg/l	Numerical		-0.209959

	TotalHardnessAsCaco3Mgl	Total Hardness As Caco3 (mg/l)	Numerical	EEA_31-01-6	-0.368179
PIP Layer	PIP_N_500_meters	Associated Nitrate Risk. - PIP Rank 1 - PIP Rank 2 - PIP Rank 3 - PIP Rank 4 - PIP Rank 5 - PIP Rank 6 - PIP Rank 7 - PIP Rank 8 = NO RISK	Qualitative / Categorical		
	PIP_P_500_meters	Associated Phosphorus Risk. - PIP Rank 1 - PIP Rank 2 - PIP Rank 3 - PIP Rank 4 - PIP Rank 5 - PIP Rank 6 - PIP Rank 7 - PIP Rank 8 = NO RISK	Qualitative / Categorical		
	PIP_P_Delivery_paths_500_meters	Associated Phosphorus Risk Classification of delivery paths in a buffer of 500 meters of the monitoring station. - Very High 1 - High 2 - Medium 3 - Low 4 - None 5 (*) Default value when no risk exists	Qualitative / Categorical		
	PIP_P_FlowPaths_500_meters	Associated Phosphorus Risk Classification of flow paths in a buffer of 500 meters of the monitoring station - Very High 1 - High 2 - Medium 3 - Low 4 - None 5 (*) Default value when no risk exists	Qualitative / Categorical		

7.9. Data storage

Chemical Data

Over 4.5 million records, precisely 4,610,145, containing chemical results were sourced from the (EPA Ireland, 2023) API. These records were methodically transformed into parquet files, each representing a distinct Irish subcatchment, and subsequently stored within the Apache Hadoop (Apache Software Foundation, 2023a) HDFS storage system. By leveraging Apache Spark's in-memory computing capabilities (Sharma et al., 2018), the data processing of this massive dataset became notably efficient. This efficiency facilitated the management of such extensive records and expedited the production of necessary queries and aggregations. The sheer volume of data and the power of in-memory computing ensured that data-driven insights could be gleaned promptly and reliably.

GIS Data

Integrating Geographic Information System (GIS) maps into the study offered a multidimensional approach to understanding the river ecosystem and its associated data. Using a combination of maps, the dataset was enriched with geographical, hydrological, and historical data, bringing a nuanced perspective to the analysis.

- Monitoring Stations
- River Water bodies cycle 2
- River Water bodies cycle 3
- Historical QValues.
- PIP layers
- Hydrometric Gauges

The map layers described above have been processed utilising geopandas and intersected with the primary dataset in the parquet file repository.

8. Correlated Chemical Parameters:

The list comprises a comprehensive set of chemical parameters that can influence water quality. These include alkalinity measures, Ammonia, oxygen levels, nitrates, nitrites, phosphate, pH, temperature, Hardness, colour, and more.

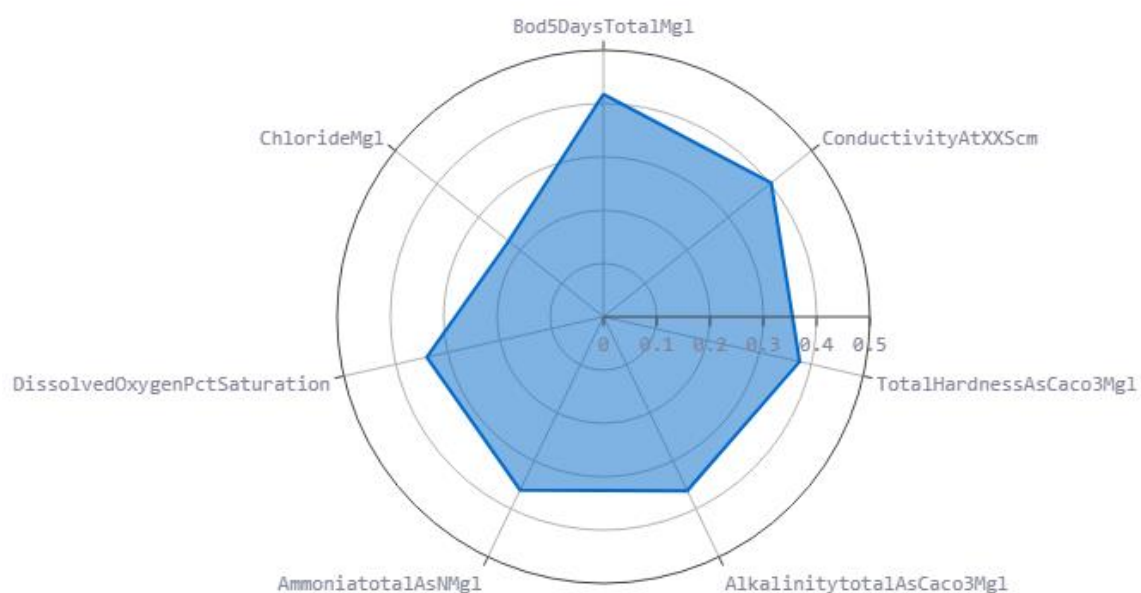


FIGURE 13: PEARSON CORRELATION OF TOP 7 PARAMETERS WITH QVALUEID

Not all of these parameters might have been discussed in the previous papers, but each plays a pivotal role in determining water quality. The final dataset includes the top seven parameters directly correlated positively or negatively with the QValueID; causation is explained to justify the selection and prioritisation of those versus the rest.

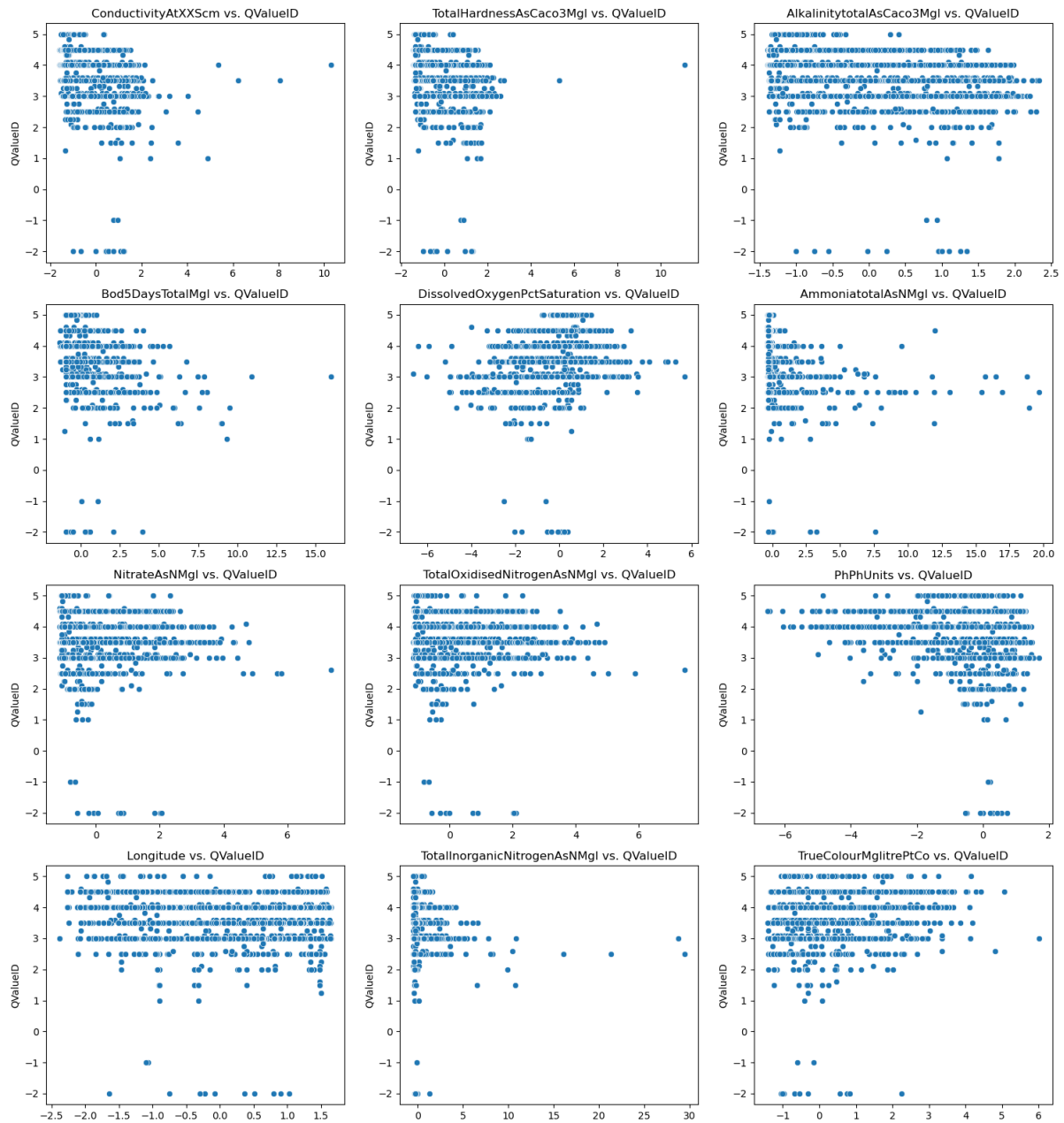


FIGURE 14: QVALUEID CORRELATION WITH MAIN PARAMETERS

8.1. Combined Conductivity @20°C and @25°C $\mu\text{S}/\text{cm}$

Conductivity at different temperatures indirectly measures the amount of dissolved salts or ions in water. These ions can come from various sources, including natural geological formations, agricultural runoff, wastewater discharges, and industrial processes. In the dataset for this study, the correlation with QValueID shows a moderate inverse correlation,

being the strongest of all other parameters at -0.4. Salinity tolerance and macroinvertebrates have been broadly studied. The authors (Dunlop et al., 2008) studied the salinity tolerance on over 100 macroinvertebrate taxa in 11 locations in Northeast Australia.

The parameter was combined since the source dataset has different measurements taken at 20°C and 25°C in different years. To preserve the integrity of this data and after consultation with biologists, a combined parameter was deemed appropriate. This decision was based on the observation that the temperature difference yielded nearly identical measurement results.

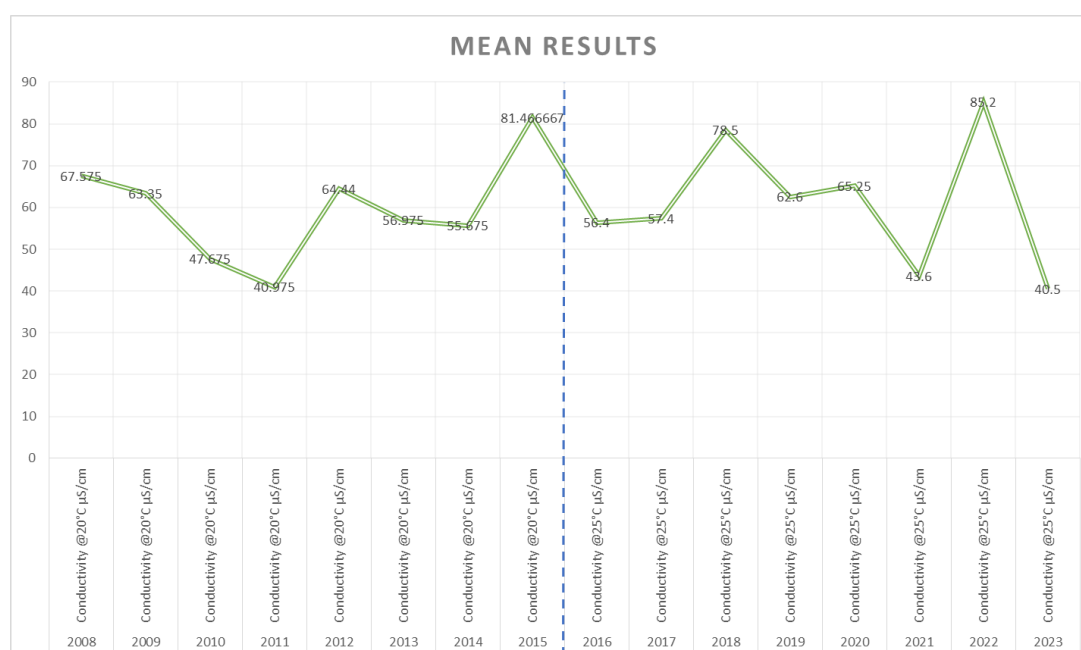


FIGURE 15: CONDUCTIVITY RS01B010100 @20°C AND @25°C (SINCE 2015) FROM EXCEL

8.2. Total Hardness (as CaCO₃) mg/l

Hard water often has a higher alkalinity, which can stabilise the pH by neutralising acids. This ensures a more stable environment, essential for many aquatic organisms. On the other hand, soft water with low Hardness may be more susceptible to pH fluctuations, which can impact aquatic life. In some cases (Kiyani et al., 2013) showing significant effects in terms of Cu and Zn toxicity in fish.

8.3. Alkalinity total As CaCO_3 Mg per litre:

This parameter shows a moderate inverse correlation with QValueID (-0.356864). Causation of the same has already been probed in several studies exploring the acidification sensitivity of macroinvertebrates. For instance, the 1987 the authors of (Allard and Moreau, 1987) observed that experimental acidification significantly reduced the overall abundance of benthic macroinvertebrates, except *Microtendipes* sp, compared to a control group.

8.4. BOD 5 Days Total Mg per litre

Biochemical Oxygen Demand (BOD) over five days, or BOD5, measures the amount of oxygen microorganisms consume in decomposing organic matter within five days. It is expressed in milligrams of oxygen consumed per litre of sample during incubation, usually at a temperature of 20°C. It is a crucial parameter for the analysis. Given its relationship with ecological status (Phu, 2014).

8.5. Ammonia total As N Mg per litre

Ecological status can be affected by high ammonia levels as it can be toxic to aquatic life, particularly to fish. Several environmental factors influence its toxicity, such as pH, water temperature, and the specific species in question.

8.6. Dissolved Oxygen % Saturation

This parameter represents the concentration of oxygen dissolved in water as a percentage of the maximum amount that could be dissolved at that temperature and atmospheric pressure. In essence, it indicates the relative amount of oxygen present compared to the total amount that could be dissolved under current conditions. Low DO can significantly affect aquatic life, increasing the risk of deterioration of the ecological status.

8.7. Nitrate (as N) mg/l

“Nitrate (as N) mg/l” represents the nitrate concentration in the water but only considering the nitrogen component of the nitrate compound. This is important for standardising measurements because nitrogen exists in various environmental forms, and focusing on the nitrogen component allows for consistent comparisons.

9.Data Preparation

Processing of the raw data set required the following transformations:

- **Remove null values; "report result":** This feature considers measurements below the detection value. So, removing null results was the only necessary processing task.
- **Yearly aggregation:** The example RDD query below is for the dataset of 6 months (~183 days) prior to the Qvalue Survey.

```
SELECT
    q.MonitoringStationCode,
    YEAR(q.DateOfSurvey) Year,
    f. parameter,
    AVG(CAST(f.ReportResult AS DOUBLE)) AS Mean_ReportResult,
    q.QValueID,
    q.QValueScore,
    q.ClassificationStatus
FROM filtered_df_view AS f
INNER JOIN qvalues_agg_df_view q ON f.MonitoringStationCode = q.MonitoringStationCode
AND f.SampleDate < q.DateOfSurvey AND datediff(q.DateOfSurvey, f.SampleDate) <= 183
GROUP BY
    q.MonitoringStationCode,
    YEAR(q.DateOfSurvey),
    f. parameter,
    q.QValueID,
    q.QValueScore,
    q.ClassificationStatus
```

- **Remove duplicate QValue Survey:**, on some occasions, for a given monitoring station, 2 or more QValues surveys were produced. In this situation, the latest completed survey of the year is considered.

```
SELECT DISTINCT data.MonitoringStationCode, data.Year, data.QValueScore, data.QValueID,
data.ClassificationStatus
FROM
(
    SELECT DISTINCT
        ROW_NUMBER() OVER (PARTITION BY StationCode, Year(DateOfSurvey) ORDER BY DateOfSurvey
DESC) as Rank,
        d.StationCode AS MonitoringStationCode,
        YEAR(d.DateOfSurvey) Year,
```

```

d.DateOfSurvey,
d.QValueScore,
d.QValueID,
d.Status AS ClassificationStatus
FROM qvalues_df_view d
) AS data
WHERE
data.Rank = 1 # Select the last qvalue survey

```

- **Generate PascalCase** parameter name. Preparing for pivoting the table.
- **Parameters first sieve:** Parameters that has been used in at least more than 4000 data points. The sieve was selected by reviewing the available dataset and balancing the appropriate number of data points that keep sufficient results from historical data for the last 20 years.
- **Combine conductivity columns:** @20 and @25 degrees as per KDR6 (Key Decision Records)
- **Remove Monitoring Stations with less than 4 years of data:** As QValues are produced every 3 years, chemical results for 4 years will only be significant to predict 1 Qvalue, insufficient to infer the rest of the target years included in the study.
- **Replace nitrate and nitrate with total nitrate.**
- **Remove parameters with insufficient samples:** SuspendedSolidsMgl only 1200 results available.
- **Filling GAPS or missing values:** For monitoring stations where there are GAPS or missing values on the series, the average of the trend series is used to replace the missing. In the example below, back or forward fill would not have been applicable.

MonitoringStationCode	Year	AlkalinitytotalAsCaco3Mgl	AmmoniatotalAsNMgl	Bod5DaysTotalMgl	ChlorideMgl	ConductivityAtXXScm	DissolvedOxygenMgl
RS06D010710	2009	200.25	0.25000	0.75	19.00	463.00	NaN
RS06D010710	2011	225.50	0.12750	1.65	17.50	329.00	NaN
RS06D010710	2015	224.00	0.14025	0.75	18.00	493.75	10.20
RS06D010710	2018	243.40	0.16780	1.36	26.08	577.00	10.00
RS06D010710	2020	226.00	0.07760	1.36	21.52	538.60	9.62

FIGURE 16: MISSING VALUES DO SERIES

The steps described above were repeated four times to generate group results in the four agreed testing periods: six months, one year, two years and all (*Refer to Key Decision Records DR3*).

For future references, we will name the dataset as dataset versions as follows:

- Dataset v0: Use parameter average aggregation of all available data.
- Dataset v1: Use parameter last six months' average aggregation data prior QValue.
- Dataset v2: Use parameter last one year average aggregation data prior QValue.
- Dataset v3: Use parameter last two years' average aggregation data prior QValue.

The query below indicates the complete set of parameter selections after the first sieve is complete. Later, correlation analysis and PCA will be discussed to provide insight into the final dataset.

```
select
  a.MonitoringStationCode,    a.Year,    a.AlkalinitytotalAsCaco3Mgl,
  a.AmmoniatotalAsNMgl,      a.Bod5DaysTotalMgl,    a.ChlorideMgl,
  a.ConductivityAtXXScm,    a.DissolvedOxygenMgl,    a.DissolvedOxygenPctSaturation,
  a.NitrateAsNMgl,          a.NitriteAsNMgl,    a.TotalInorganicNitrogenAsNMgl,
  a.OrthophosphateAsPUnspecifiedMgl,    a.PhPhUnits,    a.TemperatureC,
  a.TotalHardnessAsCaco3Mgl,    a.TotalOxidisedNitrogenAsNMgl,
  a.TrueColourHazen,        a.TrueColourMglitrePtCo,
  a.QValueID,    a.QValueScore,    a.ClassificationStatus,    b.WBCode,
  b.Geology,    b.Type,    b.SizeCat,    b.System,    b.ProtectedArea,
  b.Slope,    b.Altitude,    b.Latitude,    b.Longitude,    b.PIP_N,
  b.PIP_P,    b.PIP_P_DeliveryPoints,    b.PIP_P_FlowPaths
from data_spark_view a
inner join fulldataset b on a.MonitoringStationCode = b.MonitoringStationCode
and a.Year = b.Year
```

10. Models

10.1. Common Model Training Approach

The following definition is common to all models utilised in the study. While there might be some slightly changes in particular when using tensorflow model versus the main pyspark ones. This ensures consistency in the results and allows for an easy model comparison and deployment.

Model Preparation and Training:

- Dataset selection: v0, v1, v2, v3.
- Feature selection: PIP layers, Only Correlated Features.
- Splitting 80-20 train and test datasets

- Oversampling: River Types and Classification Status.
- Pipeline: Prepare dataset for model. String Indexer, Vector assemble, Scaled Features.
- Model definition
- Train model
 - o Fine-tuning parameter grid.
 - o Cross-validation with X folds.
- Best model performance analysis:
 - o Accuracy, F1 Score, Precision, and Recall.
 - o Save model and results.
- Plots:
 - o Model visualization: *depends on the model: decision tree or neuronal network with hidden layers*)
 - o Confusion Matrix
 - o Learning Curve.

Model training steps

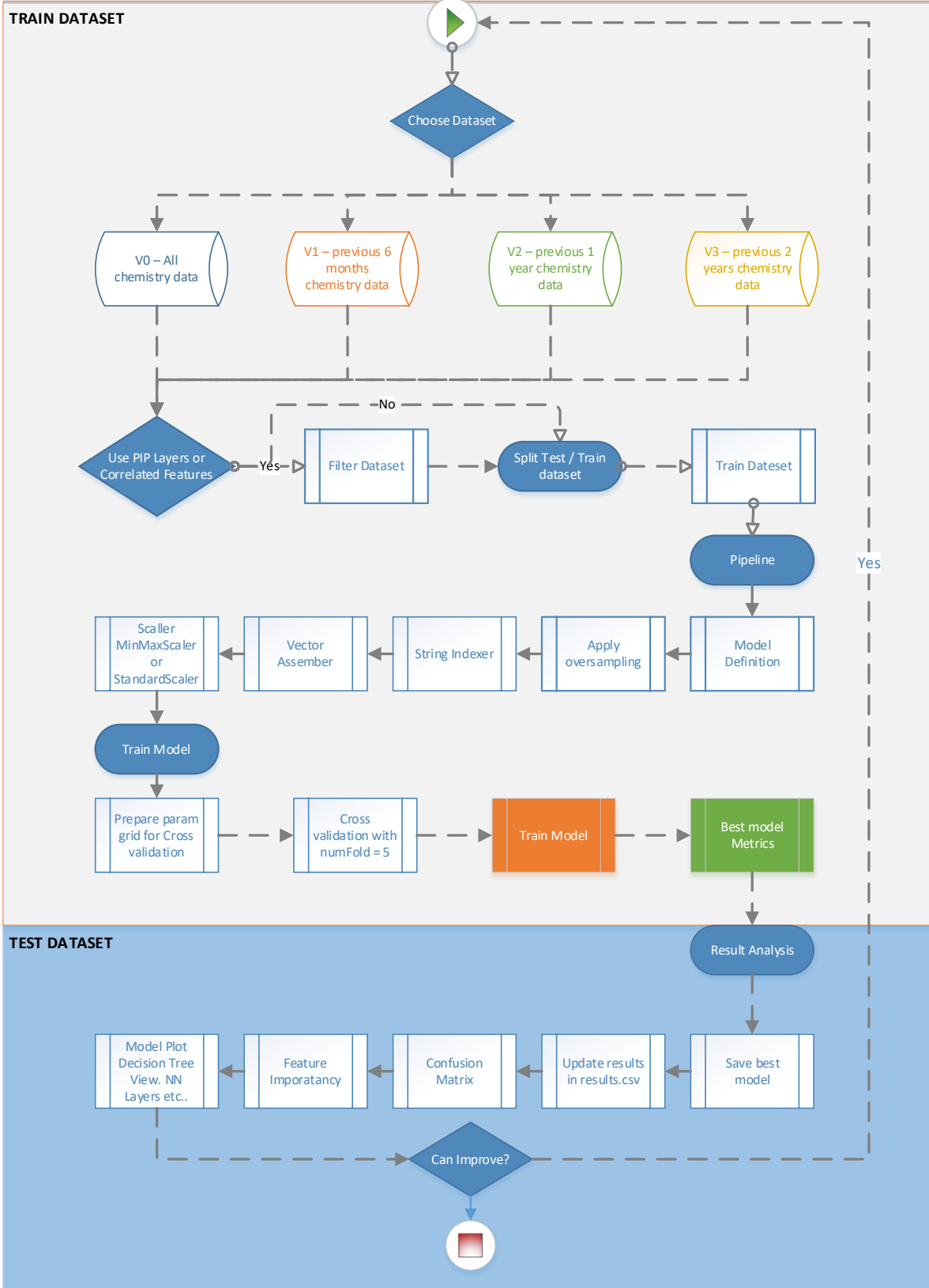


FIGURE 17: MODEL TRAINING STEPS

10.2. Naïve Bayes 44.88%

The naive Bayes classifier (Apache Software Foundation, 2023b) is simple, efficient, and can handle large feature spaces. Rooted in the principles of Bayes' theorem, this family of algorithms makes a 'naive' assumption of independence between every pair of features, thereby simplifying the computation of probabilities. This model was used first for these characteristics and to set the baseline benchmark before exploring more complex models.

Using all available features, the model performs poorly in all dataset combinations and feature selection, achieving a 44.88% accuracy on the best execution.

Results

TABLE 5: NAIVE BAYES RESULTS

Dataset	Correlated Features	Accuracy	Precision	F1 Score	Recall
V0	No	44.88%	45.59%	44.45%	44.88%
V0	Yes	44.04%	44.17%	43.59%	44.04%
V1	No	44.03%	46.47%	42.93%	44.03%
V1	Yes	41.97%	42.28%	39.66%	41.97%

10.3. Linear Support Vector Classifier 51.79%

Support Vector Machine (SVM) is a supervised machine learning algorithm for classification and regression analysis. This binary classifier optimizes the Hinge Loss using the OWLQN (Orthant-Wise Limited-memory Quasi-Newton) optimisation and L2 regularization.

The regularisation method proposed is the weighted L1-norm of the parameters for some constant C greater than 0.

$$r(x) = C||x||_1 = C \sum_i |x_i|$$

(Andrew and Gao, 2007)

Using all available features, the model performs poorly in all dataset combinations and feature selection, achieving a 51.79% accuracy on the best execution.

Results

TABLE 6: SVM RESULTS

Dataset	Correlated Features	Accuracy	Precision	F1 Score	Recall
V0	No	51.79%	49.71%	47.27%	51.79%
V0	Yes	51.42%	48.23%	46.74%	51.42%
V1	Yes	51.11%	49.60%	46.26%	51.11%
V1	No	50.45%	46.86%	46.11%	50.45%

10.4. Random Forest Classifier 83.21%

Random Forest is a supervised learning algorithm; the term "forest" denotes the construction of an ensemble of decision trees, trained using the "bagging" method.

The underlying principle of the bagging method is that amalgamating learning models amplifies the overall outcome.

Pyspark Random Forest Classifier (Apache Software Foundation, 2023c) was utilized to train the test dataset after applying a pipeline and regularization. Various iterations of the model emerged, and they excelled across different training variations and configurations of the dataset and ensembles. The configurations below were proposed:

Model 1 RF: Random Forest Classifier

This configuration engages the primary dataset with pipeline configurations integrated.

Model 2 RF-PCA: Random Forest Classifier with PCA

Similar to model 1, this configuration incorporates a dimensionality reduction via Principal Component Analysis.

Model 3 RF-ERT: Ensemble Random Forest Classifier by River Type

This ensemble approach aims to alleviate the impact of oversampling mismatched river types. It processes a subset of the test dataset based on river type, eventually combining 12 individual models.

Model 4 RF-EGT: Ensemble Random Forest Classifier by River Group

Embracing a similar objective of diminishing the reliance on oversampling, this ensemble model clusters river types based on similar sampling counts within the test dataset.

Only Models 1 and 2 are considered.

Models 3 and 4 encountered overfitting issues, especially when addressing river types with sparse sample representation. This was evident for River types 22, 23, 24, 33, and 34.

Results

TABLE 7: RANDOM FOREST RESULTS

Dataset	Model	Correlated Features	Accuracy	Precision	F1 Score	Recall
V0	RF	No	84.34%	84.22%	84.26%	84.34%
V0	RF	Yes	83.21%	83.03%	83.09%	83.21%
V2	RF	Yes	82.28%	82.06%	82.10%	82.28%
V3	RF	Yes	81.88%	81.60%	81.63%	81.88%
V0	RF-PCA	No	79.65%	79.49%	79.52%	79.65%
V0	RF-PCA	Yes	79.29%	79.08%	79.14%	79.29%
V1	RF	Yes	78.94%	78.59%	78.66%	78.94%
V3	RF-PCA	Yes	75.05%	74.71%	74.77%	75.05%
V2	RF-PCA	Yes	74.69%	74.32%	74.40%	74.69%
V1	RF-PCA	Yes	72.69%	72.61%	72.54%	72.69%

The best model configuration coded as **RF-01-v0-corr-pips** offers a good balance of excellent accuracy performance of 83.21% using the seven correlated chemical parameters for its definition. The configuration below is detailed for reproducibility purposes.

- model type: RF
- dataset version: v0
- correlated features only: true
- pip layers: true
- numTrees: 50
- numClasses=5
- numFeatures=16
- bootstrap: True
- cacheNodeIds: False
- checkpointInterval: 10

- featureSubsetStrategy: auto
- featuresCol: scaled_features
- impurity: gini
- labelCol: ClassificationStatus_index
- leafCol: default
- maxBins: **32**
- maxDepth: 30
- maxMemoryInMB: 256
- minInfoGain: 0.0
- minInstancesPerNode: 1
- minWeightFractionPerNode: 0.0
- predictionCol: **prediction**
- probabilityCol: **probability**
- rawPredictionCol: rawPrediction
- seed: -5741966080759870137 (Random)
- subsamplingRate: 1.0

The coefficient of determination R^2 , indicates how well the model's predictions match the actual outcomes; for the best model, this value is 0.7614, which means that approximately 76.14% of the variability in the dependent variable (ClassificationStatus_index) can be explained by the model's independent variables (features).

$$R^2 = 1 - \frac{\text{sum squared regression } (SS_{\text{res}})}{\text{total sum of squares } (SS_{\text{tot}})}$$

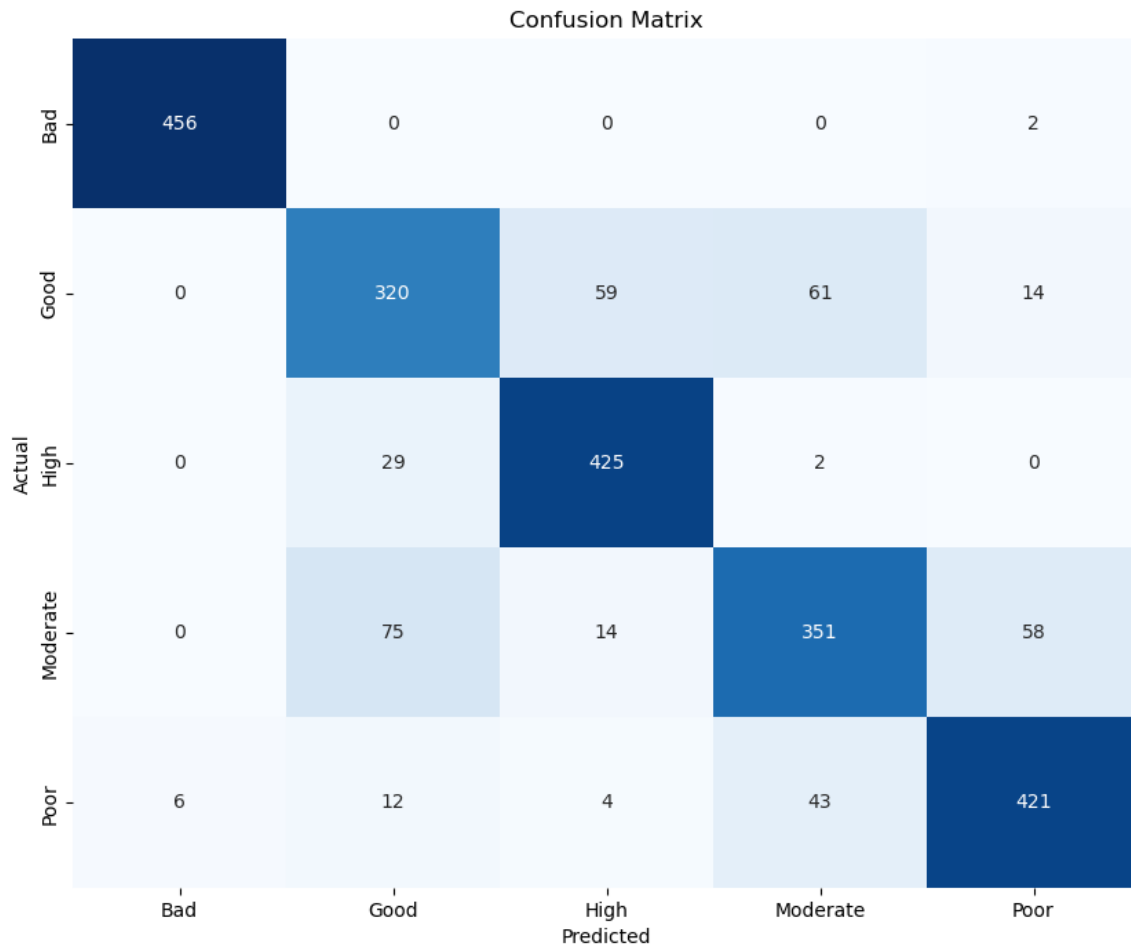


FIGURE 18: RANDOM FOREST CONFUSION MATRIX (R^2 0.7614)

Given the confusion matrix above, while the model does exhibit a misclassification rate of 17%, there is a silver lining. In more than 83% of the instances where it does misclassify, the error prediction is just one class above or below the actual class. This suggests that even when the model errs, it is not wildly off the mark but is usually close to the actual classification.

This might mean this model could still be advantageous if slight deviations from the true classification status can be tolerated for practical applications.

For the WFD objective of achieving a minimum of good classification, and considering the model's behaviour as described, a model prediction of a classification status of "Moderate", there is a high likelihood that the true classification is either "Moderate" or lies just one class above or below it. Also, it is improbable that the true status is two or more classes away from "Moderate" and the risk of it being drastically different is relatively low.

RF-01-v0-corr-pips explained:

Explaining an ensemble model consisting of 50 decision trees, a max depth of 30 nodes, five classification categories, and 16 features poses some technical challenges. As complexity and number of combinations, it quickly explodes.

For evaluation purposes, decision tree number one of **RF-01-v0-corr-pips** is presented. This single decision tree comprises 1706 tree nodes with over 121 thousand relationships created among all nodes and depths.

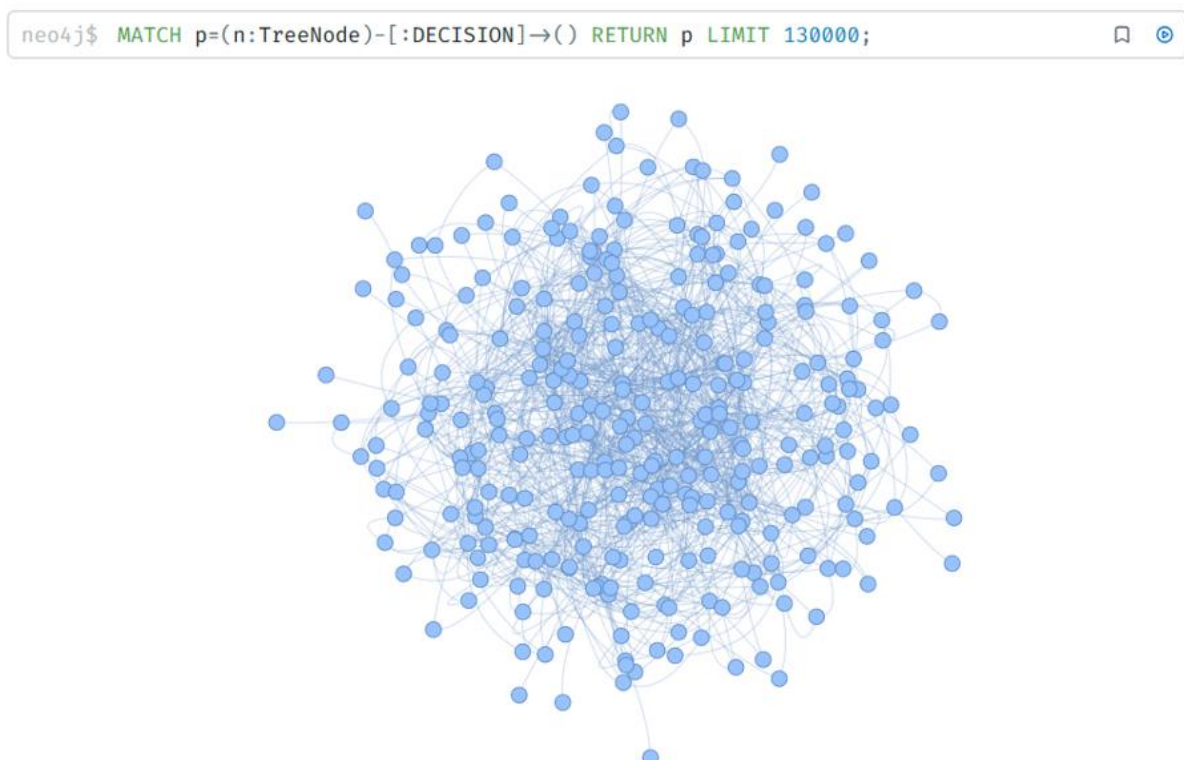


FIGURE 19: SINGLE DECISION TREE GRAPH VISUALIZATION DECISIONS

Visualisation is available under "Neo4j: Viz" tab on <https://biological-status.streamlit.app/models>

The image above was produced by constructing a graph map of the decisions and plotting the relationships of the tree nodes. Trying to present +121k decisions on the screen will not be feasible, so tools such as neo4j allow zooming and pan capabilities to be crucial. Using the graph tool, it can zoom into individual nodes and follow the different choices and decisions that a single decision model can make. To enhance clarity and understanding of each node, descriptive text outlines the feature and decision values. This text aids in discerning the direction of subsequent steps, be it to the left node or the right one..

```
(:TreeNode {info: "If (Bod5DaysTotalMg1 <= 0.04585080417456694)"})-[:DECISION {type: ">0.04585080417456694"}]->(:TreeNode {info: "If (AmmoniatotalAsNMg1 <= - 0.38364349029199574)"}))
```

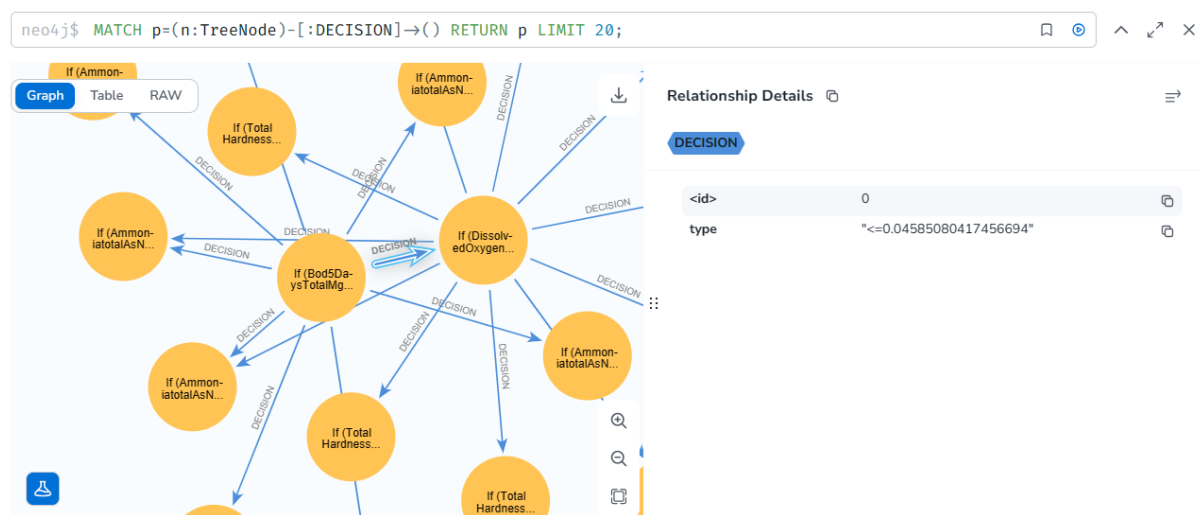


FIGURE 20: DECISION TREE GRAPH VIZ WITH ZOOM

When narrowing down the interpretation to a single prediction based on a specific input entry, tracking the models' path and outcomes becomes more feasible. It becomes pretty direct to visualize all 50 decision trees and trace only the prediction route. Here is an illustration of a prediction path leading to "poor" classification.

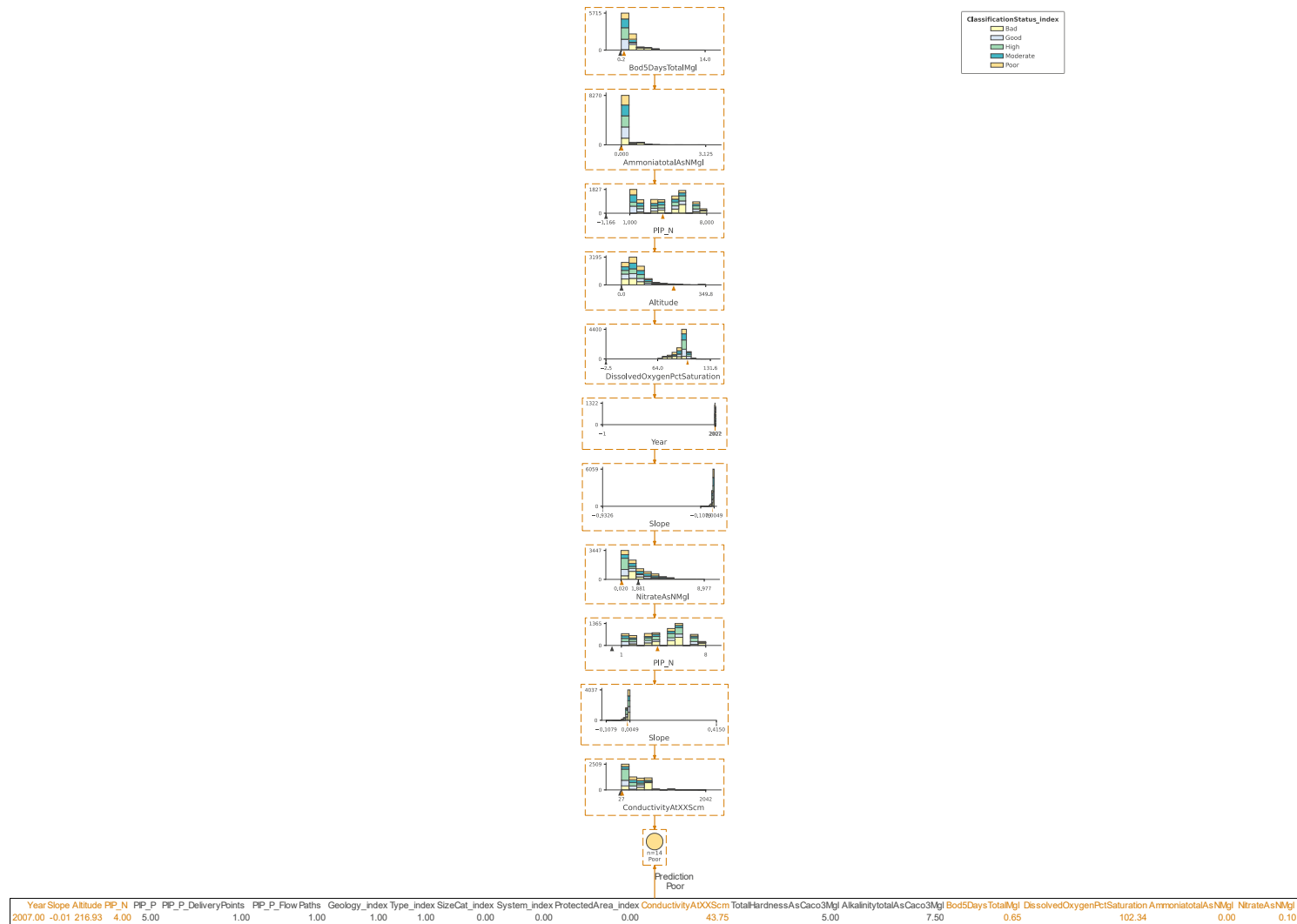


FIGURE 21: DECISION TREE PREDICTION PATH - POOR CLASSIFICATION

In this example, the decision tree prediction vote is “Poor” and has reached this conclusion after taking ten steps. BOD > Ammonia > PIP_N > Altitude > DO > Year > Slope > Conductivity.



FIGURE 22: DECISION TREE, SINGLE AMMONIA SECOND LEVEL LEAF EXAMPLE.

10.5. Multilayer perceptron 74.91%

Multilayer perceptron classifier (MLPC) is a classifier based on the feedforward artificial neural network. MLPC consists of multiple layers of nodes where each layer is fully connected to the next layer in the network with a final exit layer of 5 nodes associated with the Biological classification status. Nodes in the input layer represent the input data before preprocessing in the pipeline (standard scaler, string indexer, etc. refer to Common Model Training Approach). All other nodes map inputs to outputs by linearly combining the inputs with the node’s weights **w** and bias **b** and applying an activation function. This can be written in matrix form for MLPC with K+1 layers as follows:

$$y(x) = f_K(\dots f_2(w_{T2} f_1(w_{T1}x + b_1) + b_2) \dots + b_K)$$

On the best model, after hyperparameter tuning, the nodes in intermediate layers use ReLU function:

$$f(x) = \max(0, x)$$

Nodes in the output layer use softmax function:

$$f(z_i) = \frac{e^{z_i}}{\sum_{k=1}^N e^{z_k}}$$

The number of nodes N in the output layer corresponds to the number of classes. The image below represents the definition of the best model coded as **MLP-01-v0-corr-pips.keras**.

Model 1: MLP-01

This model uses hyperparameter tuning and cross-validation for assessing the best configuration setup. This model uses a homogeneity in the number of neurons and activation function in all hidden layers.

Model 2: MLP-02

Custom model configured for heterogeneous configuration on the number of neurons, layers and activation functions per hidden layer.

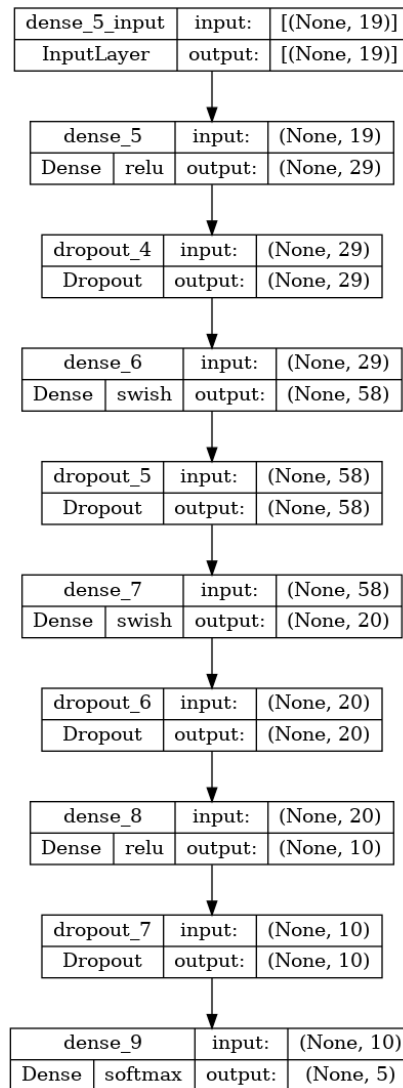


FIGURE 23: CUSTOM HETEROGENOUS NN MLP-02-V0-CORR-PIPS.KERAS

Model 02 was initially established as the benchmark for the training model. However, surpassing the prediction results of this model demanded a considerable number of iterations. The training performance deteriorated as more neurons, hidden layers, and epochs were introduced to enhance the final prediction outcomes. Interestingly, the heterogeneous setup of the model emerged as an optimal configuration for this specific Neuronal Network and the encompassing dataset. Model 01 ultimately emerged as the superior model in this context.

Further exploration into training this model using a heterogeneous approach would be beneficial. On a related note, the research by (Choudhary et al., 2023) delved into similar territory with significant success, tackling classification issues using what they termed a “Diversity Neuronal Network”. This network comprises neurons that evolve their activation functions, rapidly diversify, and then exceed the performance of homogeneous counterparts

in image classification and nonlinear regression tasks. The code for the DNN is accessible on GitHub, and a [fork](#) was utilized to process the WQI dataset for this study. Nevertheless, training the model demanded immense computational power and memory. An application was made to the Irish Centre for High-End Computing to access a supercomputer under a Class C project. However, this was not pursued further due to time limitations and response time durations.

Results:

Dataset	Correlated Features	Accuracy	Precision	F1 Score	Recall
V0	No	75.26%	74.79%	74.67%	75.26%
V0	Yes	74.91%	74.87%	74.78%	74.91%
V3	Yes	74.27%	74.48%	74.09%	74.27%
V2	Yes	73.26%	72.93%	72.89%	73.26%
V1	Yes	72.55%	72.07%	71.93%	72.55%
V0	No	72.16%	71.96%	71.82%	72.16%
V0	Yes	71.68%	71.21%	70.60%	71.68%
V2	Yes	69.25%	68.32%	68.58%	69.25%
V3	Yes	68.83%	68.23%	67.91%	68.83%
V1	Yes	68.61%	67.37%	67.68%	68.61%

MLP-01-v0-corr-pips.keras is the best model with an accuracy of 74.91% or correct predicted results, with a total of 52741 number of parameters for the training.

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	2560
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 128)	16512
dropout_1 (Dropout)	(None, 128)	0

dense_2 (Dense)	(None, 128)	16512
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 128)	16512
dropout_3 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 5)	645

=====

Total params: 52,741

Trainable params: 52,741

Non-trainable params: 0

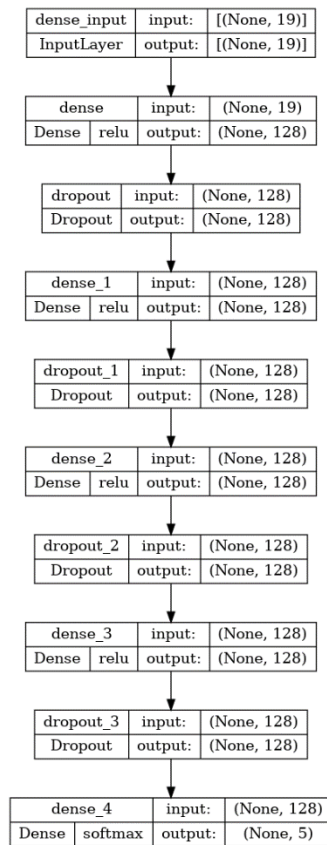


FIGURE 24: MLP-01-V0-CORR-PIPS.KERAS

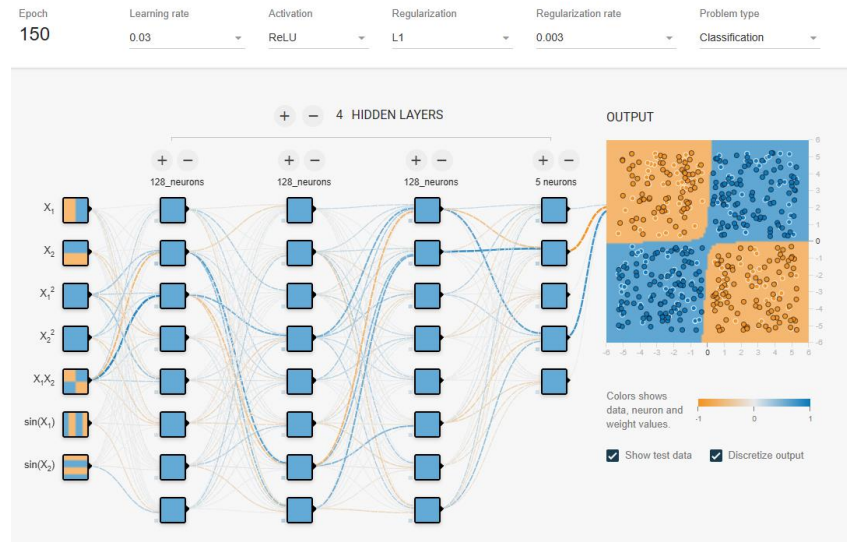


FIGURE 25: MLP-01-V0-CORR-PIPS.KERAS TENSORFLOW SIMULATION

(Carter, 2023)

A dropout regularization rate of 20% is introduced after every hidden layer to mitigate overfitting in the model. For reproducibility purposes, the following configuration indicates the setup for the best model.

- batch_size: 32
- dropout: 0.2
- epochs: 200
- hidden_layers: 4
- neurons: 128
- optimizer: adam

Confusion Matrix

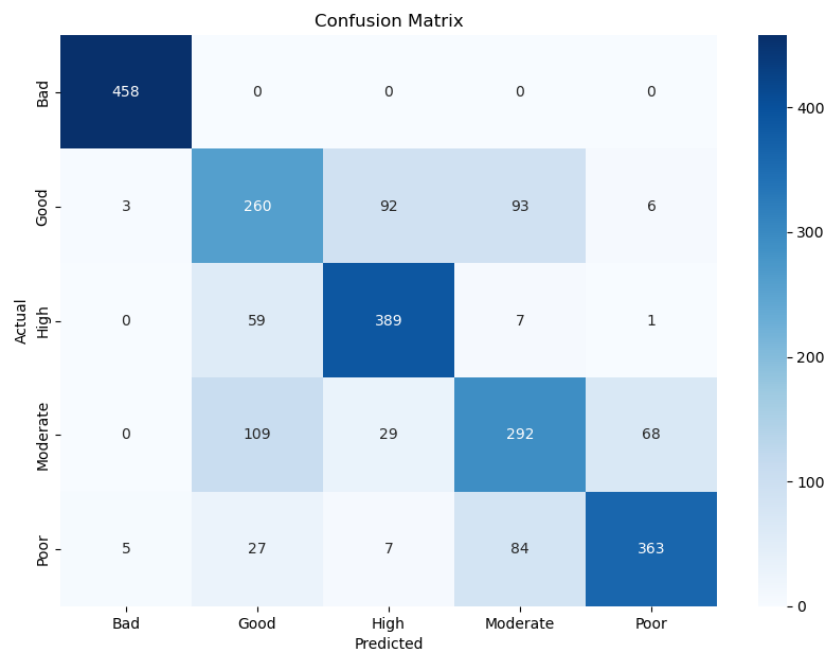


FIGURE 26: MLP CONFUSION MATRIX (BEST MODEL)

Learning curve

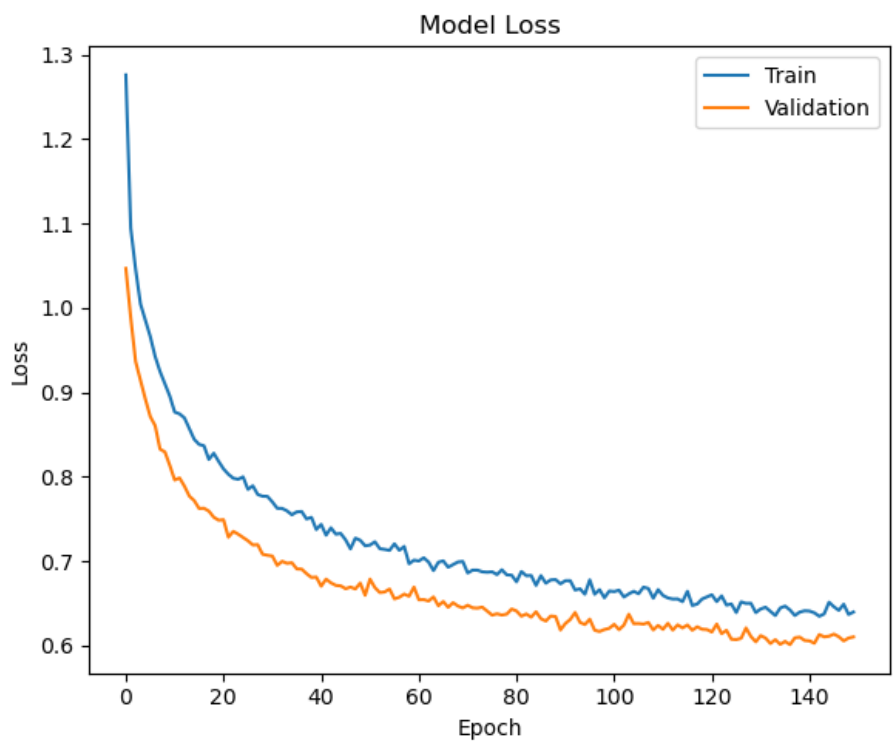


FIGURE 27: MLP LEARNING CURVE (BEST MODEL)

The Multilayer Perceptron Classifier (MLPC), based on the feedforward artificial neural network architecture, demonstrated a robust performance in classifying the Biological classification status. With its layered structure, the model utilized the ReLU activation function for intermediate nodes and the softmax function for the output layer, optimizing the classification process. Post hyperparameter tuning, the model achieved an accuracy of 74.91%, with a precision of 74.87%, an F1 score of 74.78%, and a recall rate of 74.91%. Given these performance metrics, this MLPC model presents a solid foundation for predicting Biological classification status and can be a reliable tool in relevant applications.

11. Conclusions

Analyzed the performance and characteristics of various models tested for the given problem, a few conclusions can be drawn:

Dataset	Model	Correlated Features	Accuracy	Precision	F1 Score	Recall
V0	RF	Yes	83.21%	83.03%	83.09%	83.21%
V0	MLP-01	Yes	74.91%	74.87%	74.78%	74.91%
V0	RF-02	Yes	79.29%	79.08%	79.14%	79.29%
V0	MLP-02	Yes	71.68%	71.21%	70.60%	71.68%

Adhering to a consistent model training approach provides a standardized benchmark to evaluate various models against each other. This standardization, covering aspects like dataset selection, feature handling, and model tuning, aids in ensuring consistent and comparable results across the different algorithms. While Naïve Bayes provides a good starting point due to its simplicity and efficiency, it yielded only a moderate accuracy of 44.88% at its best in this context. In the same space, the SVM, known for its optimization properties and hinge loss optimization, also resulted in moderate performance with a maximum accuracy of 51.79%. These two models were then left with their initial configuration and progressed towards a more complex model. In this case, Random Forest's Dominance was promptly revealed: Among all the algorithms tested, the Random Forest Classifier proved to be the most robust and high-performing, achieving an impressive accuracy of 83.21% in its best configuration. The strength of Random Forest lies in its ensemble methodology and ability to handle large feature spaces effectively.

Multilayer Perceptron (MLP) delivered a commendable performance, with the best model achieving an accuracy of 74.91%. The use of activation functions like ReLU and softmax in various layers showcases the non-linearity and categorical nature of the classification problem.

The high coefficient of determination (R^2) of 0.7614 for the best Random Forest model underscores its strength in predicting biological classification status. Further, its ability to often misclassify within a close range of the actual class enhances its reliability for practical

scenarios. As an extra benefit offered by the Random Forest Classifier, with just some level of sophistication and deep analysis, it can visually be described and explained following decision tree prediction routes.

Visualization tools like neo4j provide an intuitive representation of the decision-making process within the ensemble, enhancing its interpretability.

The models, particularly the Random Forest, demonstrate adaptability to different feature combinations, proving robust across different dataset variations and for the objective of achieving a precise biological classification status, the Random Forest model, particularly the RF-01-v0-corr-pips configuration, stands out as the most effective solution, balancing accuracy with interpretability, without undermining the results of the MLP that can also be considered for the same due to its significant accurate results.

Lastly, it is essential to iterate and refine these models periodically, adapting to new data and continuously improving predictive performance for the desired application.

12. Discussion and Future steps

The research journey has unveiled multiple avenues and areas that, with sufficient time and resources, warrant deeper exploration for enhanced comprehension and utility. Several areas have been left open-ended, suggesting that delving into them might lead to further advancements. The following is a detailed exploration of these areas:

Refinement of the PIP Layers Calculation:

The methodology currently in place, which calculates a 500-meter buffer around the monitoring station, reveals potential areas of refinement:

- River Flow Direction: The methodology does not consider the river's natural flow direction. Taking this into account could yield more accurate predictions and evaluations.

- Location-based Risk Assessment: Potential hazards near the monitoring station might not be correctly evaluated due to their location outside the basin or local slope and runoff factors that may exclude them from influencing the station.
- Groundwater Infiltration: Groundwater's role in a region's water quality cannot be understated. The existing calculation might overlook the complexities of groundwater infiltration, a key factor influencing water quality.

Analysis of Water Flow Patterns and Hydrometric Gauges:

An in-depth exploration of water flow patterns and the current status of hydrometric gauges in Ireland is essential. Given the direct impact of rainfall on water flow, incorporating rainfall data could prove invaluable. Although these aspects were initially set aside due to time limitations and the restricted availability of public hydrometric gauge data, experts have emphasized their significance in affecting chemical and biological statuses.

Exploration of Diversity Neuronal Networks:

Preliminary findings suggest the potential prowess of Diversity Neuronal Networks. With adequate resources, infrastructure, and time, harnessing a diversity-driven approach might yield results previously unattained, particularly in the biological status classification challenge. These networks' intricacies might offer a fresh perspective on tackling and resolving current issues.

This study has established a foundational framework, but the journey forward offers abundant opportunities for further research and enhancement. The insights derived from this research can serve as valuable references for upcoming researchers and industry specialists, facilitating well-informed decisions and spurring innovative approaches.

13. References

1. A. Danades, D. Pratama, D. Anggraini, D. Anggriani, 2016. Comparison of accuracy level K-Nearest Neighbor algorithm and Support Vector Machine algorithm in classification water quality status, in: 2016 6th International Conference on System Engineering and Technology (ICSET). Presented at the 2016 6th International Conference on System Engineering and Technology (ICSET), pp. 137–141. <https://doi.org/10.1109/ICSEngT.2016.7849638>
2. Allard, M., Moreau, G., 1987. Effects of experimental acidification on a lotic macroinvertebrate community. *Hydrobiologia* 144, 37–49. <https://doi.org/10.1007/BF00008050>
3. Andrew, G., Gao, J., 2007. Scalable training of l 1-regularized log-linear models, in: Proceedings of the 24th International Conference on Machine Learning. pp. 33–40.
4. Apache Software Foundation, 2023a. Apache Hadoop [WWW Document]. URL <https://hadoop.apache.org/> (accessed 9.21.23).
5. Apache Software Foundation, 2023b. NaiveBayes — PySpark 3.5.0 documentation [WWW Document]. URL <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.classification.NaiveBayes.html> (accessed 8.22.23).
6. Apache Software Foundation, 2023c. RandomForest — PySpark 3.5.0 documentation [WWW Document]. URL <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.tree.RandomForest.html> (accessed 9.22.23).
7. Apache Spark™ - Unified Engine for large-scale data analytics [WWW Document], n.d. URL <https://spark.apache.org/> (accessed 9.21.23).
8. Arrighi, C., Castelli, F., 2023. Prediction of ecological status of surface water bodies with supervised machine learning classifiers. *Sci. Total Environ.* 857, 159655. <https://doi.org/10.1016/j.scitotenv.2022.159655>
9. BIOMETRIC ENABLED ACCESS CONTROL, 2021.
10. Bonacina, L., Fasano, F., Mezzanotte, V., Fornaroli, R., 2023. Effects of water temperature on freshwater macroinvertebrates: a systematic review. *Biol. Rev.* 98, 191–221. <https://doi.org/10.1111/brev.12903>
11. Bullard, E., 2023. Purposive sampling. Salem Press Encycl.
12. Carter, D.S. and S., 2023. Tensorflow — Neural Network Playground [WWW Document]. URL <http://playground.tensorflow.org> (accessed 9.22.23).
13. CCT College Dublin, 2020. CCT Colleague: Data Protection Policy. URL <https://www.cct.ie/wp-content/uploads/CCTP1002-Data-Protection-Policy-2020.pdf>
14. Costa, C.J., Tiago Aparicio, J., 2020. POST-DS: A Methodology to Boost Data Science. CISTI Iber. Conf. Inf. Syst. Technol. Conferência Ibérica Sist. E Tecnol. Informação Proc. 1–6.
15. Daniels, A., Koutsougeras, C., 2021. Predicting Water Quality Parameters in Lake Pontchartrain Using Machine Learning: A Comparison on K-Nearest Neighbors, Decision Trees, and Neural Networks to Predict Water Quality, in: 2021 the 5th International Conference on Information System and Data Mining, ICISDM 2021. Association for Computing Machinery, New York, NY, USA, pp. 28–33. <https://doi.org/10.1145/3471287.3471308>
16. De Clercq, J., 2012. BitLocker in Windows 8. *Window IT Pro* 18, 93–102.
17. Department of Housing, Local Government and Heritage, 2022. Draft River Basin Management Plan for Ireland 2022-2027.







18. Derdour, A., Jodar-Abellan, A., Pardo, M.Á., Ghoneim, S.S.M., Hussein, E.E., 2022. Designing Efficient and Sustainable Predictions of Water Quality Indexes at the Regional Scale Using Machine Learning Algorithms. *Water* 20734441 14, 2801–N.PAG.
19. Developers, T., 2023. TensorFlow. <https://doi.org/10.5281/zenodo.8306789>
20. Donohue, I., McGarrigle, M.L., Mills, P., 2006. Linking catchment characteristics and water chemistry with the ecological status of Irish rivers. *Water Res.* 40, 91–98. <https://doi.org/10.1016/j.watres.2005.10.027>
21. Dunlop, J.E., Horrigan, N., McGregor, G., Kefford, B.J., Choy, S., Prasad, R., 2008. Effect of spatial variation on salinity tolerance of macroinvertebrates in Eastern Australia and implications for ecosystem protection trigger values. *Environ. Pollut.* 151, 621–630. <https://doi.org/10.1016/j.envpol.2007.03.020>
22. EPA Catchments Unit, 2021. Next generation Pollution Impact Potential maps launched. Catchments.ie. URL <https://www.catchments.ie/next-generation-pollution-impact-potential-maps-launched/> (accessed 9.21.23).
23. EPA Ireland, 2023. WFD EPA Edeniireland API [WWW Document]. URL <https://wfdapi.edeniireland.ie/docs/index#/> (accessed 7.21.23).
24. European Parliament, 2000. EU Water Framework Directive, 2000/60/EC.
25. Gómez, R., Arce, M.I., Baldwin, D.S., Dahm, C.N., 2017. Chapter 3.1 - Water Physicochemistry in Intermittent Rivers and Ephemeral Streams, in: Datry, T., Bonada, N., Boulton, A. (Eds.), *Intermittent Rivers and Ephemeral Streams*. Academic Press, pp. 109–134. <https://doi.org/10.1016/B978-0-12-803835-2.00005-X>
26. Greene, T., Shmueli, G., Ray, S., Fell, J., 2019. Adjusting to the GDPR: The Impact on Data Scientists and Behavioral Researchers. *Big Data* 7, 140–162. <https://doi.org/10.1089/big.2018.0176>
27. Horton R. K., 1965. An index number system for rating water quality. *J. Water Pollut.*
28. Islam Khan, Md.S., Islam, N., Uddin, J., Islam, S., Nasir, M.K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *J. King Saud Univ. - Comput. Inf. Sci.* 34, 4773–4781. <https://doi.org/10.1016/j.jksuci.2021.06.003>
29. Jian Sha, Xue Li, Man Zhang, Zhong-Liang Wang, 2021. Comparison of Forecasting Models for Real-Time Monitoring of Water Quality Parameters Based on Hybrid Deep Learning Neural Networks. *Water* 13, 1547–1547. <https://doi.org/10.3390/w13111547>
30. Kelly-Quinn, M., Bradley, C., Dodkins, I., Harrington, T.J., Chathain, B.N., O'Connor, M., Rippey, B., Trigg, D., 2005. Water Framework Directive: Characterisation of Reference Conditions and Testing of Typology of Rivers (2002-W-LS-7). Environmental Protection Agency.
31. Kiyani, V., Hosynzadeh, M., Ebrahimpour, M., others, 2013. Investigation acute toxicity some of heavy metals at different water hardness. *Int. J. Adv. Biol. Biomed. Res.* 1, 134–142.
32. Kotu, V., Deshpande, B., 2015. Predictive Analytics and Data Mining - Chapter 2 - Data Mining Process, in: Kotu, V., Deshpande, B. (Eds.), *Predictive Analytics and Data Mining*. Morgan Kaufmann, Boston, pp. 17–36. <https://doi.org/10.1016/B978-0-12-801460-8.00002-1>
33. Lee, E., Han, S., Kim, H., 2013. Development of Software Sensors for Determining Total Phosphorus and Total Nitrogen in Waters. *Int. J. Environ. Res. Public. Health* 10, 219–36. <https://doi.org/10.3390/ijerph10010219>
34. Leong, W.C., Bahadori, A., Zhang, J., Ahmad, Z., 2021. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). *Int. J. River Basin Manag.* 19, 149–156.
35. Li, N., Zhang, Yunlin, Shi, K., Zhang, Yibo, Sun, X., Wang, W., Qian, H., Yang, H., Niu, Y., 2023. Real-Time and Continuous Tracking of Total Phosphorus Using a Ground-Based

- Hyperspectral Proximal Sensing System. *Remote Sens.* 15. <https://doi.org/10.3390/rs15020507>
36. Li, Z., Peng, F., Niu, B., Li, G., Wu, J., Miao, Z., 2018. Water Quality Prediction Model Combining Sparse Auto-encoder and LSTM Network. 6th IFAC Conf. Bio-Robot. BIOROBOTICS 2018 51, 831–836. <https://doi.org/10.1016/j.ifacol.2018.08.091>
 37. Lu, H., Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* 249, 126169.
 38. Lusinchi, D., 2018. "The Great Fiasco" of the 1948 presidential election polls: status recognition and norms conflict in social science. *Ann. Sci.* 75, 120–144. <https://doi.org/10.1080/00033790.2018.1466194>
 39. M. Ladjal, M. Bouamar, M. Djerioui, Y. Brik, 2016. Performance evaluation of ANN and SVM multiclass models for intelligent water quality classification using Dempster-Shafer Theory, in: 2016 International Conference on Electrical and Information Technologies (ICEIT). Presented at the 2016 International Conference on Electrical and Information Technologies (ICEIT), pp. 191–196. <https://doi.org/10.1109/EITech.2016.7519588>
 40. Malinin, A., Prokhorenkova, L., Ustimenko, A., 2021. Uncertainty in Gradient Boosting via Ensembles.
 41. Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecol. Model.* 411, 108815. <https://doi.org/10.1016/j.ecolmodel.2019.108815>
 42. Mockler, E., Deakin, J., Archbold, M., Gill, L., Daly, D., Bruen, M., 2017. Sources of nitrogen and phosphorus emissions to Irish rivers and coastal waters: Estimates from a nutrient load apportionment framework. *Sci. Total Environ.* 601–602, 326–339. <https://doi.org/10.1016/j.scitotenv.2017.05.186>
 43. Mohammad Zounemat-Kermani, Youngmin Seo, Sungwon Kim, Mohammad Ali Ghorbani, Saeed Samadianfard, Shabnam Naghshara, Nam Won Kim, Vijay P. Singh, 2019. Can Decomposition Approaches Always Enhance Soft Computing Models? Predicting the Dissolved Oxygen Concentration in the St. Johns River, Florida. *Appl. Sci.* 9, 2534–2534. <https://doi.org/10.3390/app9122534>
 44. Mohammed, R., Rawashdeh, J., Abdullah, M., 2020. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. <https://doi.org/10.1109/ICICS49469.2020.239556>
 45. Nogaro, G., Mermillod-Blondin, F., 2009. Stormwater Sediment and Bioturbation Influences on Hydraulic Functioning, Biogeochemical Processes, and Pollutant Dynamics in Laboratory Infiltration Systems. *Environ. Sci. Technol.* 43, 3632–3638. <https://doi.org/10.1021/es8030787>
 46. O'Boyle, S., Trodd, W., Bradley, C., Tierney, D., Wilkes, R., S. Ní Longphuirt, J. Smith, A. Stephens, J. Barry, P. Maher, R. McGinn, E. Mockler, J. Deakin, M. Craig, M. Gurrie., 2018. Water Quality in Ireland 2013–2018.
 47. Phu, S.T.P., 2014. Research on the correlation between chlorophyll-a and organic matter BOD, COD, phosphorus, and total nitrogen in Stagnant Lake Basins. *Sustain. Living Environ. Risks* 177–191.
 48. Saunders, M., Lewis, P., Thornhill, A., n.d. *Research Methods for Business Students*, 6th ed. Pearson Education Limited.
 49. Shabani, M., Borry, P., 2018. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur. J. Hum. Genet.* 26, 149–156.
 50. Shamshirband, S., Jafari Nodoushan, E., Adolf, J.E., Abdul Manaf, A., Mosavi, A., Chau, K., 2019. Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. *Eng. Appl. Comput. Fluid Mech.* 13, 91–101. <https://doi.org/10.1080/19942060.2018.1553742>

51. Sharma, K., Marjit, U., Biswas, U., 2018. Efficiently Processing and Storing Library Linked Data using Apache Spark and Parquet. *Inf. Technol. Libr.* 37, 29–49.
52. Sheppard, V., 2020. *Research Methods for the Social Sciences: An Introduction*.
53. Struijs, J., De Zwart, D., Posthuma, L., Leuven, R.S., Huijbregts, M.A., 2011. Field sensitivity distribution of macroinvertebrates for phosphorus in inland waters. *Integr. Environ. Assess. Manag.* 7, 280–286. <https://doi.org/10.1002/ieam.141>
54. Sun, Q., Pfahringer, B., 2011. Bagging ensemble selection, in: *AI 2011: Advances in Artificial Intelligence: 24th Australasian Joint Conference, Perth, Australia, December 5–8, 2011. Proceedings 24*. Springer, pp. 251–260.
55. Tan, G., Yan, J., Gao, C., Yang, S., 2012. Prediction of water quality time series data based on least squares support vector machine. *Procedia Eng.* 31, 1194–1199.
56. Wæraas, A., 2022. Thematic Analysis: Making Values Emerge from Texts, in: Espedal, G., Jelstad Løvaas, B., Sirris, S., Wæraas, A. (Eds.), *Researching Values: Methodological Approaches for Understanding Values Work in Organisations and Leadership*. Springer International Publishing, Cham, pp. 153–170. https://doi.org/10.1007/978-3-030-90769-3_9
57. Wang, Yi, Zheng, T., Zhao, Y., Jiang, J., Wang, Yuanyuan, Guo, L., Wang, P., 2013. Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China. *Environ. Sci. Pollut. Res.* 20, 8909–8923. <https://doi.org/10.1007/s11356-013-1874-8>
58. Wei, Y., Jiao, Y., An, D., Li, D., Li, W., Wei, Q., 2019. Review of Dissolved Oxygen Detection Technology: From Laboratory Analysis to Online Intelligent Detection. *Sensors* 19. <https://doi.org/10.3390/s19183995>
59. Wienclaw, R.A., 2021. *Sampling*. Salem Press Encycl.
60. Wilkes, R., Bradley, C., Tierney, D., O’Boyle, S., Webster, P., 2018. Assessing confidence in WFD Ecological Status.
61. Yamak, P.T., Yujian, L., Gadosey, P.K., 2020. A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting, in: *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, ACAI '19*. Association for Computing Machinery, New York, NY, USA, pp. 49–55. <https://doi.org/10.1145/3377713.3377722>
62. Zhao, J., Peng, W., Ding, M., Nie, M., Huang, G., 2021. Effect of Water Chemistry, Land Use Patterns, and Geographic Distances on the Spatial Distribution of Bacterioplankton Communities in an Anthropogenically Disturbed Riverine Ecosystem. *Front. Microbiol.* 12. <https://doi.org/10.3389/fmicb.2021.633993>
63. Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., Li, L., 2021. From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? *Environ. Sci. Technol.* 55, 2357–2368. <https://doi.org/10.1021/acs.est.0c06783>
64. Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., Ye, L., 2022. A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* 1, 107–116. <https://doi.org/10.1016/j.eehl.2022.06.001>
65. Zhuang, Y., Wen, W., Ruan, S., Zhuang, F., Xia, B., Li, S., Liu, H., Du, Y., Zhang, L., 2022. Real-time measurement of total nitrogen for agricultural runoff based on multiparameter sensors and intelligent algorithms. *Water Res.* 210, 117992. <https://doi.org/10.1016/j.watres.2021.117992>

14. Annexes

14.1. Annex I: Interviews

DepthInterview-EPA	 DepthInterview-EPA.docx
DepthInterview-water utilities	 DepthInterview-WaterUtilities.docx
Participant Consent Form	 Participant Consent Form.docx
2023-08-16 17.43.07 LC- Depth Interview - WQI	 transcription-1423553601-EN.docx
2023-08-25 12.43.18 RW - Depth Interview - WQI	 Transcription-1008998402-EN.docx
2023-08-29 11.09.53 DT - Depth Interview - WQI	 Transcription-1305827341-EN.docx

14.2. Request HPC supercomputer



CCT College - Raul
Martin - HPC Nation

14.3. Chemical parameters

1,1,1,2-Tetrachloroethane µg/l	Coliform Bacteria (Total) MPN/100ml	PBDE 154 µg/l
1,1,1-Trichloroethane µg/l	Coliform Bacteria (Total) no./100mls	PBDE 28 µg/l
1,1,2,2-Tetrachloroethane µg/l	Colour Hazen	PBDE 47 µg/l

1,1,2-Trichloroethane µg/l	Colour PtCo Units	PBDE 99 µg/l
1,1-Dichloroethane µg/l	Conductivity @ 25°C (on-site) µS/cm	PCBs (Total) µg/l
1,1-Dichloroethene µg/l	Conductivity @20°C µS/cm	Pentachlorobenzene µg/l
1,1-Dichloropropene µg/l	Conductivity @25°C µS/cm	Pentachlorophenol µg/l
1,2,3-Trichlorobenzene µg/l	Copper - filtered mg/l	Perfluorooctane sulfonic acid (PFOS) µg/l
1,2,3-Trichloropropane µg/l	Copper - filtered µg/l	Perfluorooctanoic acid (PFOA) µg/l
1,2,4-Trichlorobenzene µg/l	Copper - unfiltered mg/l	Petrol Range Organics (Total) mg/l
1,2,4-Trimethylbenzene µg/l	Copper - unfiltered µg/l	Petrol Range Organics (Total) µg/l
1,2-Dibromo-3-Chloropropane µg/l	Copper - unspecified mg/l	Phenols (Total) µg/l
1,2-Dibromoethane µg/l	Copper - unspecified µg/l	Pheophytin a mg/m3
1,2-Dichlorobenzene µg/l	Cyanide (unspecified) mg/l	Picloram µg/l
1,2-Dichloroethane µg/l	Cyanide (unspecified) µg/l	Pirimiphos-methyl µg/l
1,2-Dichloroethene (Cis) µg/l	Cybutryne µg/l	Polyaromatic Hydrocarbons (PAH) -Sum µg/l
1,2-Dichloroethene (Trans) µg/l	Cypermethrin µg/l	Potassium - filtered mg/l
1,2-Dichloropropane µg/l	DDT (4-4'/P-P'isomer) µg/l	Potassium - unfiltered mg/l
1,3,5-Trimethylbenzene µg/l	Delta BHC / HCH µg/l	Potassium - unspecified mg/l
1,3-Dichlorobenzene µg/l	Depth m	Potassium IC - unspecified mg/l
1,3-Dichloropropane µg/l	Di(2-ethylhexyl) phthalate (DEHP) µg/l	Salinity 0/oo
1,3-Dichloropropene (Cis) µg/l	Diazinon µg/l	Salinity PSU
1,3-Dichloropropene (trans) µg/l	Dibromochloromethane µg/l	Salinity ppt
1,4-Dichlorobenzene µg/l	Dibromomethane µg/l	Salinity(Lab) 0/oo
2,2-Dichloropropane µg/l	Dicamba µg/l	Sample Remarks Descriptive
2,3,6,TBA µg/l	Dichlobenil µg/l	Selenium - filtered µg/l
2,4-D ng/l	Dichlorodifluoromethane µg/l	Selenium - unfiltered µg/l
2,4-D µg/l	Dichloroprop µg/l	Selenium - unspecified µg/l
2,4-DB µg/l	Dicofol µg/l	Silica (as Si) mg/l
2,6-Dichlorobenzamide µg/l	Dieldrin µg/l	Silica (as SiO2) mg/l
2-Chlorotoluene µg/l	Diesel Range Organics (Total) mg/l	Silver - unspecified µg/l
"4,4-DDD µg/l",	Diesel Range Organics (Total) µg/l	Simazine ng/l
"4,4-DDE µg/l",	Dimethoate µg/l	Simazine µg/l
4-Chlorotoluene µg/l	Dissolved Inorganic Nitrogen (as N) mg/l	Small Stream Risk Score (SSRS) Descriptive

4-Isopropyltoluene µg/l	Dissolved Organic Carbon mg/l	Sodium - filtered mg/l
4-Nonylphenol µg/l	Dissolved Oxygen % O2	Sodium - unfiltered mg/l
AMPA ng/l	Dissolved Oxygen % Saturation	Sodium - unspecified mg/l
AMPA µg/l	Dissolved Oxygen % saturation O2	Sodium IC - unspecified mg/l
Alachlor µg/l	Dissolved Oxygen mg/l	Strontium - filtered µg/l
Aldrin µg/l	Dithiocarbamates - Sum µg/l	Strontium - unfiltered µg/l
Alkalinity-total (as CaCO3) mg/l	Diuron ng/l	Strontium - unspecified µg/l
Alkalinity-total (as HCO3) mg/l	Diuron µg/l	Styrene µg/l
Alkalinity as CaCO3 – Gran titration mg/l	E. Coli MPN/100ml	Sulphate mg/l
Aluminium - filtered µg/l	E. Coli cfu/100ml	Sum 1_IWW: PBDE 28+47+99+100+153+154 µg/l
Aluminium - unfiltered mg/l	E. Coli no./100mls	Sum 3_IWW: HCHs µg/l
Aluminium - unfiltered µg/l	Endosulfan (Total) µg/l	Sum 4_IWW: Benzobfluoranthene+Benzokfluoranthene ng/l
Aluminium - unspecified mg/l	Endosulfan 1 / alpha µg/l	Sum 4_IWW: Benzobfluoranthene+Benzokfluoranthene µg/l
Aluminium - unspecified µg/l	Endosulfan 2 / beta µg/l	Sum 5_IWW: Benzog,h,iperylene+Indeno1,2,3,c,dpylene µg/l
Ammonia-Total (as N) mg/l	Endrin µg/l	Sum 6_IWW: DDT+DDD+DDE µg/l
Ammonia-Total (as NH3) mg/l	Enterococci (Intestinal) MPN/100ml	Sum 7_IWW: Aldrin, Endrin,Dieldrin, Isodrin µg/l
Ammonia-Total (as NH4) mg/l	Enterococci (Intestinal) cfu/100ml	Suspended Solids mg/l
Anthracene ng/l	Enterococci (Intestinal) no./100mls	TOC (as NPOC) mg/l
Anthracene µg/l	Epichlorohydrin (C3H5ClO) µg/l	Temperature °C
Antimony - filtered µg/l	Epoxyzonazole µg/l	Terbutryn ng/l
Antimony - unfiltered µg/l	Ethylbenzene µg/l	Terbutryn µg/l
Antimony - unspecified µg/l	Faecal coliforms cfu/100ml	Tetrachloroethene & Trichloroethene (Total) µg/l
Apparent colour Hazen	Faecal coliforms no./100mls	Tetrachloroethene µg/l
Apparent colour PtCo Units	Fats, Oils & Greases mg/l	Thallium - filtered µg/l

Appearance (on Sampling)		
Descriptive	Fenitrothion µg/l	Thallium - unfiltered µg/l
Arsenic - filtered µg/l	Flow Rate m3 per hour	Thallium - unspecified µg/l
Arsenic - unfiltered µg/l	Flow Rate m3/s	Thorium - unspecified µg/l
Arsenic - unspecified µg/l	Fluoranthene ng/l	Time sampled Descriptive
Atrazine ng/l	Fluoranthene µg/l	Tin - filtered µg/l
Atrazine µg/l	Fluoride mg/l	Tin - unspecified µg/l
BOD (6days, No inhibition) mg/l	Fluoride µg/l	Toluene µg/l
BOD (7days, No inhibition) mg/l	Gauge Reading m	Total Hardness (as Ca) mg/l
BOD - 5 days (Total) mg/l	Glyphosate ng/l	Total Hardness (as CaCO3) mg/l
BOD(2d <5°C+5d incub. 20°C) mg/l	Glyphosate µg/l	Total Nitrogen mg/l
BOD, 5 days with Inhibition (Carbonaceous BOD) mg/l	Hexachlorobenzene µg/l	Total Oxidised Nitrogen (as N) mg/l
BTX + Ethyl benzene (Sum) mg/l	Hexachlorobutadiene µg/l	Total Petroleum Hydrocarbons Descriptive
Barium - filtered µg/l	Indeno(1,2,3-c,d)pyrene ng/l	Total Petroleum Hydrocarbons mg/l
Barium - unfiltered µg/l	Indeno(1,2,3-c,d)pyrene µg/l	Total Petroleum Hydrocarbons µg/l
Barium - unspecified mg/l	Iron - filtered µg/l	Total Phosphorus (as P) mg/l
Barium - unspecified µg/l	Iron - unfiltered µg/l	Total Phosphorus (as P) µg/l
Benzene µg/l	Iron - unspecified µg/l	Total Solids mg/l
Benzo(a)pyrene ng/l	Isodrin µg/l	Transparency m
Benzo(a)pyrene µg/l	Isopropylbenzene µg/l	Tributyltin µg/l
Benzo(b)fluoranthene ng/l	Isoproturon ng/l	Trichlorobenzene (all isomers) µg/l
Benzo(b)fluoranthene µg/l	Isoproturon µg/l	Trichloroethene (all isomers) µg/l
Benzo(g,h,i)perylene ng/l	Lead - filtered µg/l	Trichlorofluoromethane µg/l
Benzo(g,h,i)perylene µg/l	Lead - unfiltered µg/l	Triclopyr µg/l
Benzo(k)fluoranthene ng/l	Lead - unspecified µg/l	Trifluralin µg/l
Benzo(k)fluoranthene µg/l	Linuron ng/l	Trihalomethanes - Total µg/l
Beryllium - filtered µg/l	Linuron µg/l	True Colour Hazen
Beryllium - unfiltered µg/l	MCPA ng/l	True Colour PtCo Units
Beryllium - unspecified µg/l	MCPA µg/l	True Colour mg/litre Pt Co
Beta-BHC /Beta-HCH µg/l	MCPB µg/l	Turbidity FTU
Bifenox µg/l	Magnesium - filtered mg/l	"Turbidity NTUs",
Boron - filtered µg/l	Magnesium - unfiltered mg/l	Unionised Ammonia - unspecified mg/l
Boron - unfiltered mg/l	Magnesium - unspecified mg/l	Uranium - filtered µg/l

Boron - unfiltered µg/l	Magnesium IC - filtered mg/l	Uranium - unfiltered µg/l
Boron - unspecified mg/l	Magnesium IC - unspecified mg/l	Uranium - unspecified µg/l
Boron - unspecified µg/l	Malathion µg/l	Vanadium - filtered µg/l
Bromobenzene µg/l	Manganese - filtered µg/l	Vanadium - unfiltered µg/l
Bromochloromethane µg/l	Manganese - unfiltered µg/l	Vanadium - unspecified µg/l
Bromodichloromethane µg/l	Manganese - unspecified mg/l	Vinyl Chloride µg/l
Bromoform µg/l	Manganese - unspecified µg/l	Visual Inspection Descriptive
Bromomethane µg/l	Mecoprop ng/l	Volatile Organic Compounds µg/l
C10-C13 Chloroalkanes µg/l	Mecoprop µg/l	Xylenes (Total) µg/l
COD-Cr mg/l	Mercury - filtered µg/l	Zinc - filtered mg/l
Cadmium - filtered µg/l	Mercury - unfiltered µg/l	Zinc - filtered µg/l
Cadmium - unfiltered µg/l	Mercury - unspecified µg/l	Zinc - unfiltered mg/l
Cadmium - unspecified µg/l	Methylene Chloride / Dichloromethane µg/l	Zinc - unfiltered µg/l
Calcium - filtered mg/l	Mineral oils µg/l	Zinc - unspecified mg/l
Calcium - unfiltered mg/l	Molybdenum - filtered µg/l	Zinc - unspecified µg/l
Calcium - unspecified mg/l	Molybdenum - unfiltered µg/l	alpha BHC / Alpha-HCH µg/l
Calcium Hardness (as CaCO3) mg/l	Molybdenum - unspecified µg/l	alpha-Hexabromocyclododecane (HBCDD) µg/l
Calcium IC - filtered mg/l	Naphthalene µg/l	beta-Hexabromocyclododecane (HBCDD) µg/l
Calcium IC - unspecified mg/l	Nickel - filtered µg/l	gamma-BHC / HCH (Lindane) µg/l
Carbon Tetrachloride µg/l	Nickel - unfiltered µg/l	gamma-Hexabromocyclododecane µg/l
Chlorfenvinphos µg/l	Nickel - unspecified µg/l	meta + para-Xylene µg/l
Chloride mg/l	Nitrate (as N) mg/l	n-Butylbenzene µg/l
Chlorobenzene µg/l	Nitrate (as NO3) mg/l	n-Propylbenzene µg/l
Chloroform µg/l	Nitrite (as N) mg/l	o,p-DDT µg/l
Chloromethane µg/l	Nitrite (as N) µg/l	o,p-TDE µg/l
Chlorophyll mg/m3	Nitrite (as NO2) mg/l	ortho-Phosphate (as P) - unspecified mg/l
Chlorophyll µg/l	Nonylphenol ethoxylates (Sum) µg/l	ortho-Phosphate (as P) - unspecified µg/l
Chlorpyriphos Ethyl µg/l	Nonylphenol-diethoxylate µg/l	ortho-Phosphate (as PO4) mg/l
Chlorpyriphos µg/l	Nonylphenol-hexaethoxylate µg/l	ortho-Xylene µg/l

Chromium - filtered µg/l	Nonylphenol-monoethoxylate µg/l	pH (on-site) pH units
Chromium - unfiltered µg/l	Nonylphenol-pentaethoxylate µg/l	pH measured at: °C
Chromium - unspecified µg/l	Nonylphenol-tetraethoxylate µg/l	pH pH units
Clpyralid µg/l	Nonylphenol-triethoxylate µg/l	para-tert-Octylphenol µg/l
Cobalt - filtered µg/l	Orthophosphate (as P) - filtered mg/l	sec-Butylbenzene µg/l
Cobalt - unfiltered µg/l	PBDE 100 µg/l	tert-Butylbenzene µg/l
Cobalt - unspecified µg/l	PBDE 153 µg/l	

14.4. Model Results (csv)

This file includes all results and model trains, which include overfitting models and other models not finally considered.



results.csv

Abbreviations

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Averages
Cl	chloride
CNN	Convolutional Neural Network
CODMn	Permanganate Index
DO	Dissolved Oxygen
DWT	Discrete Wavelet Transform
LSTM	Long Short-Term Memory
MAE	mean absolute error
NH3-N	Ammonia Nitrogen
NO3	nitrate
Nox	nitrogen oxides
PCA	Principal Component Analysis
pH	potential of hydrogen

RF	Random Forest
RMSE	root mean square error
SVM	Support Vector Machine
SVR	Support Vector Regression
TDS	total dissolved solids
TN	Total Nitrogen
TP	total phosphorus
TP	total phosphorus
WQI	Water Quality Index
WT	Water Temperature