**Project Proposal**

# **Introduction**

For this project, we selected the [Water Quality Dataset](https://www.kaggle.com/datasets/sukhmandeepsinghbrar/water-quality) [1] (Washington State, USA) available on Kaggle. The dataset contains 1,048,575 records spanning from 1970 to 2011, providing detailed measurements of multiple chemical and environmental indicators collected from different water bodies across Washington. These indicators include parameters such as pH, dissolved oxygen, turbidity, nitrate levels, and temperature, among others.

The dataset is structured in tabular form, where each row represents a specific measurement instance, and each column corresponds to a specific parameter or metadata feature. Its long-time span and large volume make it well-suited for machine learning (ML) models, which thrives on large, complex datasets and can uncover patterns not easily discovered by traditional statical methods.

Water quality datasets of this scale are valuable for predictive modelling as it reflects seasonal variations, environmental changes, and long-term trends across different locations. These factors are central to evaluating the Water Quality Index (WQI), an indicator for assessing suitability of water for human consumption, agriculture, and ecosystem health.

Given increasing concerns surrounding climate change, long-term water quality trends become more relevant than ever. By selecting this dataset, ML methods can be applied to predict the WQI while exploring the relative influence of both chemical indicators and seasonal/temperature-based environmental factors over time.

# **Literature Review**

Machine learning has become an increasingly important approach in water quality research, particularly for predicting the Water Quality Index (WQI). Numerous studies have shown that ML models can achieve high predictive accuracy, although gaps remain in terms of feature explainability, environmental factors, and scalability.

In *“[Analyzing Water Quality Datasets, Key Parameters and Trends](https://download.ssrn.com/2024/8/12/4923373.pdf?response-content-disposition=inline&X-Amz-Security-Token=IQoJb3JpZ2luX2VjEJf%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIQCD7VrROgNALcOVIvSqaSvOLqwiHObzcgNMF153zFoJhQIgJY2ME%2BiVqGtPmUxm2bM8YpYaPUdUyBpUFTX7muyjCF0qvgUIHxAEGgwzMDg0NzUzMDEyNTciDGxTtf1LXdQVEqW9LiqbBQe%2Fpj9kCLGcfr%2F7fe4QFiNf5Rl5mH%2BpWPzPkeCeCPa0Kp0SVkqoDqA6tLNh7PjMH1ysgU%2BpIhm69xBIBXjfI7B%2FounnoIq6dJk7ZRKGOWAvoJ2t%2Fzi2NdPiVJkwkqxeorjSi6fUv6z59ZXjQnh1ZQNi7Wubk2uyMKEwTHYa%2FuCRGJ90aV%2FQ5KhlBuUb9fUNEMQODb5bV5NtbzLc1EBz%2FkOhB%2FEVoqbrVQFpLbmjHUhXieqcGUn5lMrUbWgKVT9zSiOFAVFAbD0AGr0xPNDzbNnmKRomsIUGRH6ZESvmc3SRlMHfcet%2FPbEbkpYRLyXE85p2bo2Ud1X88%2FkEjCj8aQveh5R74RJ08Ewj8mC%2FVe%2Fa7zTQPRZRdW086Gw7mEnomwujxefYZQu1aUwr%2B8Fpea7pNQV0Od%2FOnvzWal96imkZhBM4lCB4W6bf1uLP72sAlryQstutu%2B5Flb%2B8yOd020ZWp4LpEHZeWNA8mP%2BbL7DVm1CGU6O6Pxwp0tXlDSCsKH8pHOmJ5a2KEp0OySvduMqA0ru9wE9gQo%2FgRkSJbk8JdkE0rNA3DO43UawM1wOI%2BsAR6rOuhLfU1Nj%2BcmiTZ5NB7m86apVCqpUojEynrPIRG37vynLHfCdjmhOn2iPnV2uLGBA9Lg2XIX%2FNNC9mpm2NbcdENZMc3R35FADGwQ%2B8d%2BSi6KC4rtTU5nwA6ukVa40fmRJSH%2Bqq2oB0rTtWFHyhe7AM4laXdKs6fkpsD8r9ZHgxMdSYAsMBKrZQNq5NWyN9xboh%2B2hzXL6VsZcP6sLCiEkKAJizVeAM38UrPIIl5YGLbS9jatwFsoAYS9tJ61f0MHkfDO3HRzo5R9pRdx6ej5GMJ9qLZyi3AqK9VTKje95SJO7KscAPLa4whO3BxgY6sQHUgOoik39oWTqMffS8hbFZD7ctAp5VgrhMr91mSzA1rgv15ADP9x7Z4l%2FNEncCFOvjGwZ8bNGp8I8B4kknZfIu%2BdQ7hfR8qj%2BRcFbc1oap5CP107xxqf%2BcVu5OcIdFocmb29morlqPIDwU0nMp0VHQ9d0egKQ0%2B%2Bj9g2%2F8%2FrrW3LVEXADL9WHAmA2n2i53OH9zmvzcAmgA5i1qbue7rmPIyJvolStFxNCtjJ3%2FbUkY30k%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20250921T224922Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAUPUUPRWE22W3XV4J%2F20250921%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=516e974587236254db19b013ca912fea9a7fdad8ad43889c059da7ab31095514&abstractId=4923373)”*[2]a large-scale survey of water quality datasets hosted on Kaggle was performed. They screened more than 100 collections and identified the most common physical, chemical, and biological parameters. Their analysis emphasised that ensemble methods such as Random Forests, XGBoost, and LightGBM consistently outperformed simpler models like logistic regression and KNN. The study also highlighted the growing use of deep learning for capturing complex non-linear patterns. However, the review focused largely on model performance and dataset availability, with limited attention to how environmental variables such as temperature or seasonality influence outcomes.

In *“*[*Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms*](https://link.springer.com/article/10.2991/hcis.k.211203.001)*”*[3], a range of algorithms including Random Forests, Support Vector Machines, Multinomial Logistic Regression, Neural Networks, and Bagged Trees were applied to an Indian dataset. Their models reached very high levels of accuracy, in some cases above 99%. The key drivers of prediction were dissolved oxygen, nitrate, total coliform and pH. Despite these strong results, the dataset was relatively small, and the study concentrated mainly on classification of WQI categories, without considering long-term or seasonal dynamics.

In *“*[*Machine Learning Models for Water Quality Assessment and Prediction*](https://ieeexplore.ieee.org/abstract/document/10842914/metrics#metrics)*”*[4]seven well-established algorithms-Naïve Bayes, Random Forest, Decision Trees, AdaBoost, Logistic Regression, K-Nearest Neighbours, and SVM were tested on 21,000 water quality samples. Their findings showed Random Forest to be the best performing model with an accuracy of 91%, while also providing insights into feature importance. Once again, dissolved oxygen and pH emerged as the most influential parameters. Although this study provided a solid comparison of methods, it did not address the impact of wider climatic or temporal drivers of water quality

In “[Water Quality Prediction Using KNN Imputer and Multilayer Perceptron](https://www.mdpi.com/2073-4441/14/17/2592#B27-water-14-02592)” [5] deep learning architectures, particularly multilayer perceptrons studied modelling water potability. The study employed a KNN imputer to handle missing data and compared neural network performance with classical ML approaches. Results demonstrated that deep learning could successfully model water quality, often reaching comparable accuracy to ensemble methods. However, the dataset used was relatively small (935 samples) and restricted to binary potability classification, with no integration of broader environmental or seasonal factors. This limited the generalisability of the findings, though the study confirmed the potential of neural networks in this field.

In “[Seasonal Changes in Water Quality and its Main Influencing Factors in the Dan River Basin](https://www.sciencedirect.com/science/article/pii/S0341816218304417)” [6] the authors examined spatial and seasonal variations in surface water quality over a six-year period in China. Their analysis showed that nitrate nitrogen (NN) and total phosphorus (TP) were the main pollutants, with levels fluctuating significantly across the seasons. They also found that vegetation cover (NDVI), patch connectivity, and hydrological factors such as water level played important roles, though the influence of these factors shifted with the time of year. For example, vegetation cover was most influential in spring and summer, while land-use factors became more dominant in winter. The study demonstrated that water quality is shaped by both natural and human driven processes, and that these drivers vary with the seasons. However, the work relied on statistical methods rather than predictive machine learning models, and it did not test how temperature and seasonal variables compare directly with chemical indicators in forecasting WQI. This gap strengthens the case for our project, which applies machine learning to a large dataset to explore not just the chemical drivers of water quality, but also the role of seasonal and climatic influences.

In “[Seasonal Variations and Assessment of Surface Water Quality Using Water Quality Index (WQI) and Principal Component Analysis (PCA): A Case Study](https://www.mdpi.com/2071-1050/16/13/5644)” [7] the authors investigated the Nador Canal in Morocco to assess how water quality changes across the year. Samples were collected from 22 sites in both summer and winter and analysed using WQI and PCA. The study found that water quality was notably poorer in summer, with higher temperatures and lower dissolved oxygen levels reducing suitability, while winter showed some improvement. However, heavy metal and nutrient pollution from agricultural and industrial runoff became more evident in the colder months. These results underline the strong influence of rainfall, land use and seasonal conditions on water quality. Although the study demonstrated the importance of seasonal drivers, it relied solely on statistical techniques rather than predictive machine learning. It did not attempt to quantify how seasonal factors compare with chemical indicators in determining WQI. This gap is one that our project addresses by applying ensemble and explainable ML approaches to a much larger dataset, enabling us to evaluate both chemical and environmental drivers of water quality together.

In “[Spatial and temporal variations in the relationship between lake water surface temperatures and water quality – A case study of Dianchi Lake](https://www.sciencedirect.com/science/article/pii/S0048969717335453)” [8] authors examined how lake surface water temperature (LSWT) shapes ecological conditions and contributes to water quality changes. Using twelve years of monitoring data (2005-2016) from Dianchi Lake, they developed a hybrid forecasting approach that combined support vector regression (SVR), principal component analysis (PCA), and back-propagation artificial neural networks (BPANN). Their results showed that LSWT is a fundamental driver of eutrophication, with thresholds between 17.6–18.5 °C triggering increases in chlorophyll-a, chemical oxygen demand, and total nitrogen. Spatial analyses further indicated that pollution patterns spread from the northern to the southern parts of the lake, with the north remaining the most degraded. The study highlights how rising LSWT, in combination with nutrient loading and urbanisation, amplifies the risk of cyanobacterial blooms. However, its scope was largely confined to temperature quality interactions within a single lake system, leaving broader hydrological and land use dynamics less explored.

In “[Impacts of Climate Change on Groundwater Quality: A Systematic Literature Review of Analytical Models and Machine Learning Techniques](https://iopscience.iop.org/article/10.1088/1748-9326/adb8ff/meta)” [9] provided a wide-ranging overview of how climate change is shaping groundwater systems. The review drew attention to the role of rising temperatures, shifting rainfall patterns, and extreme events such as floods and droughts in accelerating groundwater contamination and salinisation. It also compared different modelling strategies, noting that analytical tools such as the Analytical Hierarchy Process (AHP), GIS-based DRASTIC models, and MODFLOW remain common for assessing vulnerability, while machine learning techniques including Random Forest, SVM, and XGBoost are increasingly used to predict groundwater quality parameters and contamination risks. A key point made by the authors is that hybrid approaches, which bring together physical process-based models and machine learning, often provide the most reliable forecasts. Nonetheless, they identified several weaknesses in the current literature, particularly the lack of high-resolution datasets, limited consideration of extreme climate events, and weak integration between climatic, hydrological, and chemical factors. These findings strengthen the case for studies that move beyond purely chemical predictors and explicitly test the influence of climatic and seasonal variables on water quality outcomes.

In “[Seasonal Dynamics of Water Quality in Response to Land Use Changes in the Chi and Mun River Basins, Thailand](https://www.nature.com/articles/s41598-025-91820-4#:~:text=Key%20findings%20include:%20(1),expansion%20contributed%20to%20its%20deterioration)” [10] land use shifts, especially the growth of agriculture and urban areas, were examined and had affected water quality over a fourteen-year period (2007-2021). They tracked eleven water quality indicators and applied redundancy analysis (RDA) to study seasonal and spatial changes. The results showed clear seasonal differences: dissolved oxygen and electrical conductivity were typically higher in the dry months, while the wet season brought spikes in pollutants such as nitrates, ammonia, coliform bacteria and suspended solids due to runoff. Land use changes intensified these effects, with farming and urbanisation worsening water quality, while forests and aquatic areas helped to filter and reduce pollutants. The study underlined the strong influence of human activity and seasonal rainfall on water systems but mainly used statistical analysis rather than predictive machine learning. This leaves space for future research to test how seasonal and land use factors compare with chemical indicators in machine learning models.

Overall, the reviewed studies indicate that both ensemble methods and deep learning models can produce strong results in WQI prediction. Yet, several limitations are consistently observed. Most existing research relies on relatively small datasets, often limited to a few thousand records rather than large-scale collections. In addition, seasonal and environmental factors are rarely incorporated, despite their clear influence on water systems. Finally, there has been little emphasis on explainable ML techniques, leaving uncertainty around the relative importance of different predictors. The present project seeks to address these gaps by applying ensemble and explainable ML approaches to a dataset of more than one million water quality records from Washington State. The aim is not only to enhance predictive accuracy, but also to provide insight into how seasonal variation and temperature interact with chemical indicators in determining water quality.

# **The Research Question**

*Can ensemble and explainable machine learning methods assign Water Quality Index (WQI) scores and predict long-term changes caused by seasonal and temperature influences, revealing a correlation between climate change and water quality?*

Water Quality Index (WQI) is a widely used tool for classifying water quality based on several parameters which may include, for example, pH level, dissolved oxygen, phosphate level, nitrate level, suspended solids and temperature. Once the factors are chosen, each factor is assigned a weight, where the total weight of all aggregated factors is equal to 1. The WQI provides an interpretable score based on multiple parameters, which allows for policymakers, researchers, and the public to easily assess the water quality. [11] By applying ensemble ML methods, we aim to assign reliable WQI scores to each water body recorded in the dataset.

Human-related activities, such as agriculture, urbanisation and mass industrial processes, have deteriorated the quality of water by releasing harmful chemicals such as phosphates and nitrates into waterways. Excess nutrients from fertilizers and manure can leach into water, causing algae blooms and as a result, a depleted dissolved oxygen level. [12] Human activities have also contributed to a warming planet, with average temperature increasing steadily over the past number of decades [13]. Thermal warming of rivers and lakes can have a detrimental impact on freshwater ecosystems by exacerbating problems of eutrophication, causing immense harm to aquatic wildlife.[14]

As our chosen dataset spans across several decades, it is possible to analyse long-term trends in both chemical pollutants and climatic variables. This makes it possible to determine the extent of how temperature and seasonal factors influence water quality compared to chemical indicators, revealing the impact climate change may have on WQI scores. By pinpointing the dominant parameters causing changes in water quality over a long period of time, policymakers can take appropriate action to implement safeguards to minimise the effects of certain detrimental factors on the quality of water.

**References**

[1] ‘Water Quality’. Accessed: Sept. 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/sukhmandeepsinghbrar/water-quality

[2] A. Jawed, W. Den, and I. Alsmadi, ‘Analyzing Water Quality Datasets, Key Parameters and Trends’, Aug. 06, 2024, *Social Science Research Network, Rochester, NY*: 4923373. doi: 10.2139/ssrn.4923373.

[3] Md. M. Hassan *et al.*, ‘Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms’, *Hum-Cent Intell Syst*, vol. 1, no. 3, pp. 86–97, Dec. 2021, doi: 10.2991/hcis.k.211203.001.

[4] S. N. Raj and M. Maanasa, ‘Machine Learning Models for Water Quality Assessment and Prediction’, in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, Nov. 2024, pp. 1–6. doi: 10.1109/IDICAIEI61867.2024.10842914.

[5] ‘Water Quality Prediction Using KNN Imputer and Multilayer Perceptron’. Accessed: Sept. 22, 2025. [Online]. Available: https://www.mdpi.com/2073-4441/14/17/2592#B27-water-14-02592

[6] G. Xu *et al.*, ‘Seasonal changes in water quality and its main influencing factors in the Dan River basin’, *CATENA*, vol. 173, pp. 131–140, Feb. 2019, doi: 10.1016/j.catena.2018.10.014.

[7] ‘Seasonal Variations and Assessment of Surface Water Quality Using Water Quality Index (WQI) and Principal Component Analysis (PCA): A Case Study’. Accessed: Sept. 22, 2025. [Online]. Available: https://www.mdpi.com/2071-1050/16/13/5644

[8] K. Yang, Z. Yu, Y. Luo, Y. Yang, L. Zhao, and X. Zhou, ‘Spatial and temporal variations in the relationship between lake water surface temperatures and water quality - A case study of Dianchi Lake’, *Science of The Total Environment*, vol. 624, pp. 859–871, May 2018, doi: 10.1016/j.scitotenv.2017.12.119.

[9] T. N. Chowdhury, A. Battamo, R. Nag, I. Zekker, and M. Salauddin, ‘Impacts of climate change on groundwater quality: a systematic literature review of analytical models and machine learning techniques’, *Environ. Res. Lett.*, vol. 20, no. 3, p. 033003, Mar. 2025, doi: 10.1088/1748-9326/adb8ff.

[10] ‘Seasonal dynamics of water quality in response to land use changes in the Chi and Mun River Basins Thailand | Scientific Reports’. Accessed: Sept. 22, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-91820-4

[11] Md. G. Uddin, S. Nash, and A. I. Olbert, ‘A review of water quality index models and their use for assessing surface water quality’, *Ecological Indicators*, vol. 122, p. 107218, Mar. 2021, doi: 10.1016/j.ecolind.2020.107218.

[12] T. M. Addiscott, *Nitrate, Agriculture and the Environment*. CABI, 2005.

[13] J. Hansen, M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade, ‘Global temperature change’, *Proceedings of the National Academy of Sciences*, vol. 103, no. 39, pp. 14288–14293, Sept. 2006, doi: 10.1073/pnas.0606291103.

[14] M. F. Johnson *et al.*, ‘Rising water temperature in rivers: Ecological impacts and future resilience’, *WIREs Water*, vol. 11, no. 4, p. e1724, 2024, doi: 10.1002/wat2.1724.