ADVANCING WATER QUALITY PREDICTION THROUGH INTEGRATING MACHINE LEARNING WITH DATA AUGMENTATION: A CASE STUDY FOR FIRST NATIONS COMMUNITIES IN BRITISH COLUMBIA

by

Anqi Chen

B.Sc., Wenzhou University, 2021

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
NATURAL RESOURCES AND ENVIRONMENTAL STUDIES

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

April 2024

Abstract

Clean drinking water access is essential for public health and regarded as a scarce resource for Indigenous communities in rural and remote areas. In this research, a new iron and manganese prediction method based on Data Augmentation and Machine Learning Algorithms to be applied to drinking water in BC's First Nation communities is reported. GAN based modelling and NI-BS-NI based modelling were developed to investigate the effects of different data augmentation methods and predictors for iron and manganese prediction results. Reliable synthetic data was obtained through both data augmentation methods, allowing 4 machine learning algorithms to predict iron and manganese utilizing 3 and 5 physical properties respectively. Compared with RF, XGB, and DT machine learning models, the GBR model showed the strongest fitting ability and accurate predictions for both NI-BS-NI based modelling and GAN based modelling in predicting iron and manganese, with the Train R^2 and Test R^2 of two models nearing 1, and all the RMSE scores are below 0.06. The decision-making tool developed using GAN technology is considered to have greater application potential due to its ability to provide accurate predictions while requiring only 3 input physical parameters.

Table of Contents

List of Tables

List of Figures

Glossary

ANN	Artificial neural network
AO	Aesthetic objective
BC	British Columbia
BP-ANN	Backpropagation Artificial Neural Networks
BPNN	Back Propagation Neural Network
DA	Data augmentation
DT	Decision Tree
FNs	First Nations
GAN	Generative Adversarial Network
GBR	Gradient Boosting Regression
GRNN	Generalized Regression Neural Networks
KNN	K-Nearest Neighbors
MAE	Mean absolute error
ML	Machine learning
MLR	Multiple Linear Regression
NI-BS-NI	Numerical Interpolation-Bootstrapping-Noise Injection
NN	Neural Network
PM	Particulate matter

R^2	Coefficient of determination
RF	Random Forest
RMSE	Root mean square error
SHAP	SHapley Additive exPlanations
SRR	Small, Rural, and Remote
SVM	Support Vector Machine
SWDNs	Small water distribution networks
TDS	Total dissolved solids
WQC	Water quality class
WQI	Water quality index
XGB	Extreme Gradient Boosting

Acknowledgement

First and foremost, I am profoundly thankful to my supervisor, Dr. Jianbing Li, for his invaluable guidance and feedback throughout my master's research at the University of Northern British Columbia. He is a scholar of great academic ability, meanwhile, he is the most patient and responsible mentor. Sincerely thanks to Dr. Li again.

I extend my sincere appreciation to my co-supervisor, Dr. Min Zhao, and my committee members, Dr. Oliver Iorhemen and Dave Tamblyn, for their valuable feedback and constructive criticism of this thesis. Their expertise and guidance have enriched the quality of my research greatly.

I am grateful to the members of my research group, Min Xie, Sorour Nasimi, and Mostafa Dorosti, for their collaboration and support. I am lucky to have the opportunity to work alongside such talented individuals. I would like to extend a special thank you to Cheng Lu, whose mentorship and friendship have been a source of inspiration and support.

Lastly, I am deeply grateful to my friends and my best team members, Dixuan Li, Jianliang Mao, Qing Guo, Wanhua Shen, Wei Deng for their love, encouragement, and support. To my mom and dad, thank you for your endless encouragement and for always believing in me. Their simple yet powerful message—that happiness is everything—has kept me grounded and focused on what truly matters.

Chapter 1 Introduction

Ensuring the availability of safe drinking water is an essential concern for both public health and overall development (World Health Organization, 2023). The attainment of universal access to secure water and sanitation infrastructure remains an ongoing challenge, primarily attributable to historical differentials and the marginalization of distinct demographic cohorts (Brown et al., 2023). The current scarcity of drinking water resources is mainly manifested in the contamination of drinking water sources, uneven distribution of water resources, over exploitation of groundwater, inadequate sanitation facilities, and susceptibility to extreme weather events and changes in water circulation patterns (Patrick et al., 2019).

Many Indigenous communities live in remote regions, resulting in comparatively greater challenges in accessing clean drinking water than residents in urban areas (Balasooriya et al., 2023). For example, there exists a significant disparity for Indigenous households in Canada. The likelihood of lacking access to clean drinking water is 90 times higher than non-Indigenous households (Wolfe, 2006; Balasooriya et al., 2023). These vulnerabilities do not stem from insufficient capacity or lack of interest within the communities, but rather as an outcome of the structural frameworks inherited from the colonial state (Baijius & Patrick, 2019; Wolfe, 2006).

The quality of drinking water is a critical determinant of public health, and the presence of metals can significantly influence its overall safety and potability (Stride et al., 2023). Indigenous communities emerge as particularly susceptible to exposure to toxic metals (Balasooriya et al., 2023; Navarro-Espinoza et al., 2021). It is worth mentioning that trace elements in water, such as iron (Fe), and manganese (Mn), are essential nutrients

required to maintain human metabolism in appropriate amounts. They are crucial for physiological metabolic processes of human activities and the human nervous system, and an excess or deficiency can potentially lead to health issue (Le Bot et al., 2016; Zoni & Lucchini, 2013).

The methods commonly employed for quantifying metal concentrations in water predominantly rely on sophisticated laboratory instruments, such as Atomic Absorption Spectroscopy (AAS) and Inductively Coupled Plasma Mass Spectrometry (ICP-MS). Consequently, there are limitations imposed by experimental conditions, substantial instrument costs, and time consumption (Hu et al., 2019).

Machine learning (ML) is a very powerful method for data analysis. Numerous models have been developed using ML algorithms for the analysis of water quality and water security issues (Azrour et al., 2022). ML is expected to serve as a viable alternative to traditional water sampling, especially in measuring challenging-to-assess water quality parameters (Chowdhury et al., 2009; Shahi et al., 2020). Water samples from Small, rural, and remote (SRR) communities necessitate transportation to analytical laboratories for the measurement of metal concentrations, entailing considerable investments in both time and financial resources (Mian et al., 2020). Hence, the search for an effective and resource-efficient method to monitor the drinking water quality in First Nations SRR communities is a pressing concern and holds significance in addressing environmental injustices resulting from political and historical factors (Wolfe, 2006).

In this study, the efficiency of diverse ML tree models was systematically examined for predicting metal concentrations based on water quality detection data collected from 2019 to 2020. The employed ML algorithm was elucidated, followed by a comprehensive

explanation of the model prediction outcomes to understand the impact of various predictors on water quality fluctuations. The subsequent sections provided a thorough discussion of the results for comparative analysis. There are two objectives in this study. Firstly, it aims to forecast the levels of iron and manganese utilizing data derived from primary indicators employed in water quality assessment. Through hyperparameter optimization, the ML model's parameters are fine-tuned and optimized to enhance predictive accuracy. Secondly, the study developed a graphical user interface employing optimal features and ML algorithms within the Python programming framework to facilitate the prediction of iron and manganese content at sampling sites.

Chapter 2 Literature Review

2.1 Overview of Canada's First Nations

2.1.1 History and current situation

2.1.1.1 History

First Nations, Inuit, and Métis collectively fall under the term "Aboriginal" in Canada and are referred to as "Indigenous" globally (Woodcock, 1988). Typically, the term "First Nations" in Canada referred to communities residing south of the tree line, predominantly situated below the Arctic Circle (Assembly of First Nations, 2021). First Nations in Canada, historically referred to as "Indians," in contemporary discourse, many groups prefer the term "First Nations" as a more accurate and respectful alternative to "Indians" (Indigenous Foundations, 2009). Members of First Nations typically identify with their specific nation, such as Mohawk, Cree, Oneida, and others, emphasizing their unique cultural affiliations (Government of Canada, 2017). Unique social and cultural communities are formed by Indigenous Peoples. These areas have experienced historical displacements because of the colonial expansion in Europe (Kingsbury, 1998). They differ from the current dominant society in terms of cultural, economic, and political characteristics due to their distinct cultural and traditional knowledge that shapes their connections with the environment and the society (United Nations, 2023). First Nations communities in Canada have similarities to the Indigenous people of America and Australia with similar historical colonial backgrounds (Daley et al., 2018; Rowles III et al., 2020). The historical legacies of colonialism and exclusion have led to widespread

challenges in poor governance and obstacles to resources access (Brown et al., 2023), including access to clean water which is a basic human right (United Nations, 2015).

2.1.1.2 Population and distribution

In 2021, the Indigenous population in Canada, numbering 1.8 million as enumerated during the census. This figure significantly surpasses both the count of First Nations people residing in Australia and New Zealand (Statistics Canada, 2022).

Presently, Canada acknowledges 617 officially recognized First Nations governments or bands, with approximately 50% distributed in the provinces of Ontario and British Columbia (BC) (Figure 2.1) (Government of Canada, 2017).

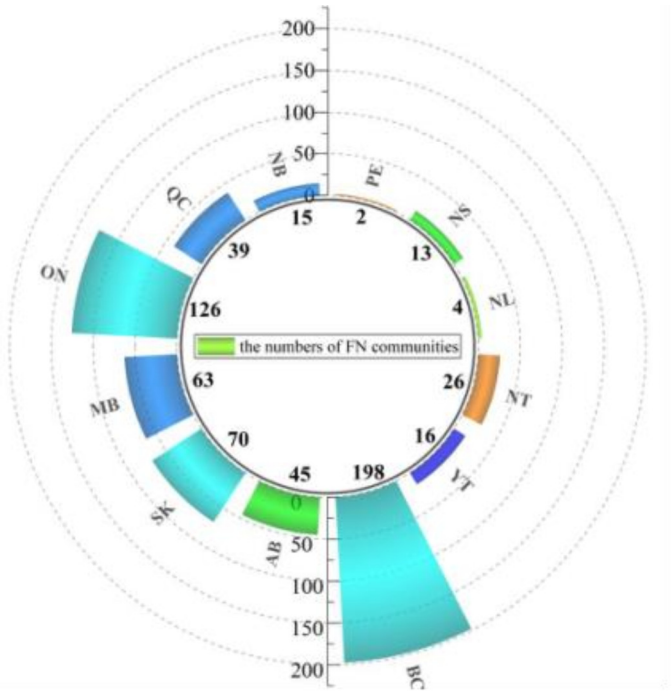


Figure 2.1 Distribution of FNs communities in Canada

2.1.2 First Nations communities in BC

This study utilizes water quality data from five different FN communities distributed across various regions in BC, Canada to predict metal content in First Nations regions, especially SRR communities in BC by using ML models. FN communities are typically located in small rural and remote areas, characterized by dispersed rural living patterns (Baijius & Patrick, 2019).

2.1.3 Drinking water quality and security

High-income countries like Canada and the United States exhibit differences in universal water access, primarily stemming from the scale and geographical distribution of drinking water, issues related to racial wealth disparities, identity, and institutionalized marginalization structures (Meehan et al., 2020). The dispersed rural living patterns of FNs often results in inadequate infrastructure, aging water treatment equipment, and distance from urban areas collectively contributing to the challenges in accessing clean drinking water. A significant number of individuals in FNs communities, especially SRR communities, lack access to an adequate supply of quality tap water in their households even though tap water is the main source of drinking water in Canada. However, the crisis of water insecurity faced by Indigenous families in Canada remains insufficiently investigated (Duignan et al., 2022). As of October 25, 2021, long-term drinking water advisories persist in 31 communities, affecting 43 small water systems on First Nations reserves (Government of Canada, 2023). In British Columbia, 19 SRR First Nations communities face three water quality advisories, eight boil water advisories (mainly due to E. coli contamination), and ten "do-not-consume" advisories as of September 2021 (McLeod et al., 2020; First Nations Health Authority, 2023). Several first nation remote

communities in British Columbia, Canada, have experienced issues such as poor aesthetic properties, the presence of coliforms, and elevated concentrations of metals (Hu et al., 2022).

A common water infrastructure widely used in many First Nations communities is a concrete-constructed household water cistern. A truck driving through the community from the water treatment plant provides water to each household's water tank weekly (McLeod et al., 2014). Aging and undisinfected tanks and trucks can cause possible contamination of drinking water. Winter freezing and thawing events lead to concrete household water tank damages, allowing pollutants like organic matter and rodents to infiltrate, causing drinking water insecurity, thereby compromising the safety of drinking water (Baijius & Patrick, 2019). Currently, residents of LTFN rely on bottled water, which is replenished every two weeks (Islam & Yuan, 2018; Pang et al., 2021). In a press conference held on November 19, 2021, the LTFN emphasized the urgent need for federal funding to ensure access to clean drinking water, stating, "We need to have water which is safe. There is no alternative." (Prince George Citizen, 2021).

This predicament not only necessitates an escalation in governmental investment in water infrastructure but also imposes a substantial financial burden on public health protection (Li et al., 2021).

2.2 Heavy metals: Iron and Manganese in Drinking Water

Ensuring the safety of drinking water is essential as it serves as the foundation for human survival, ecological well-being, and agricultural systems. It is among the most important factors in guaranteeing proper functioning of human society (Schimpf & Cude,

2020). Common water pollutants can generally be classified based on the nature of pollutants into categories such as: organic pollutants, inorganic pollutants, and microbial pollutants (Martin & Johnson, 2012). Organic pollutants encompass various organic compounds originating from agriculture, industrial emissions, and urban sewage, including dissolved organic matter, fats, proteins, organic solvents, as well as volatile organic pollutants (VOCs) from processes like chemical manufacturing, petroleum refining, and printing. Microbial pollutants involve bacteria, viruses, and parasites from sewage, livestock farming, and agricultural runoff, potentially leading to the spread of waterborne diseases. One of the most significant inorganic pollutants is heavy metals, and certain heavy metals can cause serious harm, such as mercury, lead, cadmium, copper, and nickel, along with their compounds.

2.2.1 Sources of iron and manganese

2.2.1.1 Natural factors

Iron and Manganese are abundant in nature, existing naturally in the water supply due to catchment and erosion.

Iron is the fourth most abundant element in the Earth's crust (Sun et al., 2023). Iron in the Earth's crust enters underground water and groundwater through the following three routes:

a. The oxide of divalent iron in the rock layer is converted into soluble iron by the groundwater containing carbonation.

b. The oxide of tri-iron is reduced to divalent iron and then is dissolved in the underground water body by carbonic acid.

c. There is a large amount of organic matter in the underground environment, such as organic acids, which can dissolve iron.

The source of manganese and its distribution in the environment are also very extensive (Teng et al., 2001), being present in almost all rocks. Manganese is released from the native minerals after weathering and combining with some oxygen-containing ions or molecules to form secondary minerals. Under the condition that the soil environment is acidic (PH < 5.5), these secondary minerals are dissolved into soluble manganese and some of them will penetrate the water (Zhai et al., 2021).

2.2.1.2 Human action factors

However, with the rapid development of global industry, the accumulation and cycling of heavy metal elements in ecosystems have been induced by anthropogenic interventions. Substantial pollution hazards to the environment have resulted from the emissions of heavy metals in industrial processes, including waste gases, wastewater, and waste residues generated in various industrial activities namely ore extraction, alloy smelting, leather manufacturing, electroplating, battery production, plastic manufacturing, ceramic firing, paper printing, fossil fuel combustion, and chemical textile processes (Lim & Aris, 2014; R. Singh et al., 2011; Yeganeh et al., 2023). After the untreated discharge of such waste, heavy metal emissions in the atmosphere can settle into water bodies through precipitation and atmospheric deposition. The waste residue deposited in the soil also pollutes the surface water and groundwater through surface runoff, soil erosion, seepage and infiltration. Consequently, heavy metals continue to accumulate in aquatic ecosystems (Ayangbenro & Babalola, 2017). In certain situations, iron, as a metal used in the

manufacturing of pipes and faucet components within water supply systems, may be released into tap water (Veschetti et al., 2010).

2.2.2 Impact of iron and manganese

2.2.2.1 Human health concerns

Heavy-metal-induced water pollution poses severe environmental challenges and hazards to the entire ecosystem. As persistent and toxic pollutants, heavy metals，can be transmitted through the food chain into human bodies. Proteins and biocatalysts in human tissues can react with heavy metal ions entering the body, leading to their aggregation and structural changes that result in loss of activity. Meanwhile, heavy metal ions continue to accumulate in the human body until their concentration reaches or exceeds the detoxification threshold of human organs (Briffa et al., 2020). This accumulation leads to pathological changes in human organs, causing acute or chronic poisoning, even carcinogenic. Additionally, they can migrate into animals and human bodies through respiration, contact, and other pathways, exerting irreversible toxic effects and causing functional damage (Gumpu et al., 2015b). Compared to other pollutants, heavy metals in water exhibit more noticeable latency, toxicity, and recalcitrance. They can invade the human body directly or indirectly by drinking and skin infiltration, accumulating in organs such as the kidneys and liver (Jaishankar et al., 2014; Singh et al., 2023; Zhang et al., 2023).

Iron and manganese are indispensable elements in human physiological metabolism. However, the excessive concentration of these two elements will lead to human metabolic disorders and induce various diseases (Gumpu et al., 2015a; Valko et al., 2005). The maximum concentration for manganese allowed in drinking water is 0.12mg/L and 0.3 mg/L for iron, based on Guidelines for Canadian Drinking Water Quality (Health

Canada, 2019). The standard of drinking water in the world clearly stipulates the content of iron and manganese: the sum of iron content is 0.3 mg/L, and the allowable concentration of manganese is 0.1 mg/L (World Health Organization, 2017).

Persistent intake of water with excessive iron content can lead to chronic poisoning. Symptoms include significant iron deposits in the liver and spleen, and may also result in osteoporosis, cirrhosis, coronary heart disease, diabetes, and reduced insulin secretion, thereby causing disruptions in carbohydrate metabolism in the human body.

Manganese exhibits higher toxicity in its divalent state compared to trivalent manganese, potentially leading to conditions such as tremor paralysis, memory decline, and pneumonia. Elevated manganese levels can also have adverse effects on the central nervous system, initially manifesting as neurasthenia and dysfunction in the autonomic nervous system and potentially development of Parkinson's syndrome in the later stages, along with certain impacts on reproductive capacity and cognitive functions (Kim et al., 2022). Data from surveys suggests that workers in manganese mines are susceptible to severe mental disorders resembling schizophrenia. Additionally, cases of illness and fatalities have been reported among residents in the outskirts of Tokyo, Japan, who consumed well water contaminated with manganese.

2.2.2.2 Impact on drinking water

Metals such as lead, arsenic, copper, and chromium can find their way into drinking water sources through geological processes, industrial discharges, or aging infrastructure. Their presence, even in trace amounts, can have profound effects on human health and ecosystems (Lu et al., 2015). The main concern of iron and manganese in drinking water is their effects on drinking water taste, odor and color (Schwartz et al., 2021). The Canadian

drinking water guidelines state that manganese in drinking water requires supervision according to health risks and aesthetic considerations (Health Canada, 2021). Aesthetic objective (AO) or recommended value for Fe and Mn is specified in water quality guidelines (Hu et al., 2022). The amount of manganese will directly affect the chromaticity of the water. If the iron concentration in water surpasses 0.3 mg/L, the water turns cloudy, and when it is more than 1 mg/L, the water develops an iron-like taste. When the content of manganese in the water is above 0. 5mg/L, the water produces a special odor and an unpleasant color. The occurrence of phenomena like "red water" and "black water" is attributed to water with elevated levels of manganese and iron. According to CBC News, in Cape Breton FN reserve, the excess Mn and Fe concentrations caused the aesthetic objectives for drinking water stated in the Canadian guidelines were substandard and received the "do-not-consume" advisory. Members in Cape Breton FN reserve seriously protested the dark, odor tap water. Additionally, the specific conditions under which manganese causes coloration can vary, including factors such as pH, oxygen levels, and the presence of other minerals in the water. From a sensory perspective, washing clothes and utensils with water containing high levels of iron and manganese can easily lead to discoloration, affecting functionality and aesthetics (Meena et al., 2005). When iron and manganese accumulate significantly in water supply pipelines, the transport capacity of the pipeline is significantly reduced due to blocked water pipes (Tremblay et al., 1998).

A study indicated 4% of the surveyed FNs households across Canada showed manganese concentrations in stagnant (first draw) or flushed tap water that exceeded the health-based maximum acceptable concentration (MAC) defined by the 2019 Guidelines for Canadian Drinking Water Quality (Health Canada, 2019). The fact is 12.8% of

households had manganese concentrations higher than the AO in their flushed tap water, in addition, 3.5% of households had iron levels over the AO (Schwartz et al., 2021). A survey was done in metropolitan France questioning families with children aged 6 months up to 6 years. The results showed that the concentration of Mn and Fe in the tap water was so high that at bare minimum the readings exceeded at least one of the highest-level regulations set by regulatory authorities (Le Bot et al., 2016). Lheidli T'enneh First Nation community (Prince George, British Columbia) expressed apprehension regarding the presence of excessive iron and manganese in their drinking water. The concern stems from the consistent failure of the treatment systems to meet the manganese or hardness treatment objectives, even after the modification of the existing equipment settings in August 2021, which still did not lead to proper treatment.

2.2.3 Current detection approaches

2.2.3.1 Field sampling and analysis

This conventional method for directly measuring metal concentrations in water samples involves on-site collection of water samples, followed by analysis using laboratory instruments. At present, the laboratory instrument analysis methods mainly include Atomic Fluorescence Spectrometry (AFS) (Fernández-Martínez et al., 2015), Atomic Absorption Spectrometry (AAS) (Bua et al., 2016), Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES) (Zhao et al., 2015) and Inductively Coupled Plasma Mass Spectrometer (ICP-MS) (Deng et al., 2018). These methods can quantify heavy metals accurately, but the pretreatment processes are inconvenient, costly, and require a wide range of operational expertise.

2.2.3.2 Sensor technologies

Sensor technology involves sensors and monitor devices for monitoring water quality on site and recording real-time water data. Due to the strength of this method, such as immediacy, portability, and high selectivity for specific metals, sensor technology has been widely used in the field of identifying metal ions and detecting the variation of metal concentration. However, several challenges obstruct promoting the applications, including the expense of instrumentation and the maintenance of the sensor for the reason to ensure the sensitivity and reliability.

Microbial electrochemical sensors characterize the concentration of heavy metals by exploiting the property of a decline in electrochemical activity in bacteria under the influence of heavy metals, resulting in a degradation of their output electrical signals (Wang et al., 2020). Electrochemical sensors based on metal-organic frames (MOF) can achieve robust, sensitive, selective and reliable sensing of metal ions (Shafqat et al., 2023). Nanomaterials-based chemical sensors are widely employed as effective analytical tools for the detection of heavy metal ions. They exhibit characteristics such as high sensitivity, portability, overall optimized detection capability and performance (Alias et al., 2020; Rasheed et al., 2022). In addition to simple and reliable electrochemical methods, there are spectroscopic and optical methods applied for sensing of metal ions (Harrington et al., 2011).

2.2.3.3 Modelling and simulation

Multiple models, such as hydrogeological or mathematical models, combined with sampled data and Geographic Information Systems (GIS), have been utilized to simulate metal distribution and predict the metals concentration in water bodies (Motovilov &

Fashchevskaya, 2021). In academic field, integrating various modelling strategies has been commonly introduced to acquire comprehensive and reliable results. Hydrodynamic models simulate hydrodynamic processes such as water flow, dissolution phenomena, sedimentation, and the transport of suspended particles to infer the transport and distribution patterns of heavy metals in water. Artificial intelligence models leverage techniques such as ML and deep learning to recognize complex patterns within water quality data, facilitating the prediction of heavy metal concentrations in water. Back propagation neural network (BPNN) was applied in heavy metal concentration prediction in the Qinghai-Tibet Plateau basin. The model predicted the content of 4 heavy metals (As、Sb、Mo、Mn)，while pH, dissolved oxygen (DO), conductivity (EC), total phosphorus (TP) and iron (Fe) were used as the input values (Xiao et al., 2023).

ML models have become a hot topic in recent years for predicting water quality, although, there is a dearth of research specifically addressing the prediction of heavy metal concentrations. The introduction to ML is developed in detail in the next section.

2.3 Machine learning overview

2.3.1 Models introduction

This study presents a new framework based on data augmentation algorithm, which combines recent water quality data from SRR First Nation areas in BC and 4 ML models, and the importance information of relevant features can be effectively captured, thus improving prediction accuracy. The following are the main concepts and components of ML:

a. Training Data: The training process of a ML model relies on extensive data. This data contains information relevant to the task the model needs to learn.

b. Features: Features are critical attributes that describe the data. In ML, selecting and extracting appropriate features is crucial for the performance of the model.

c. Model: A model is a mathematical representation used to capture patterns within the data. The choice of the model can be varied, such as classification, regression, or clustering.

d. Training: During the training phase, the model learns patterns and relationships from the training data. This typically involves adjusting the model's parameters to accurately represent the data.

e. Testing and Validation: After training, the model needs to be validated using test data. This helps assess the model's generalization ability.

f. Prediction and Decision: Once training is complete, the model can be utilized to make decisions on given data. This is the goal of ML.

The following are presented to the four ML models used in this study.

2.3.1.1 Random Forest (RF)

Random Forest (RF), as depicted by Figure 2.2, employs decision trees as sub-classifiers through an ensemble learning approach, combining multiple decision trees to form the Random Forest. In Random Forest, the original dataset is partitioned into multiple subsets, and each decision tree' sub-classifier employs a distinct method of optimal attribute splitting. This ensures that each tree's training process yields different results, guaranteeing their distinctiveness.

It is described as an improvement upon Bagging (Bootstrap Aggregating), with a key distinction lying in the introduction of random feature selection. During selecting split points for each decision tree, RF randomly chooses a subset of features and then performs traditional split point selection on this subset. In addition, compared to Bagging, RF shows rapid training process and better generalization ability, benefiting from its flexibility feature (Lin et al., 2017).

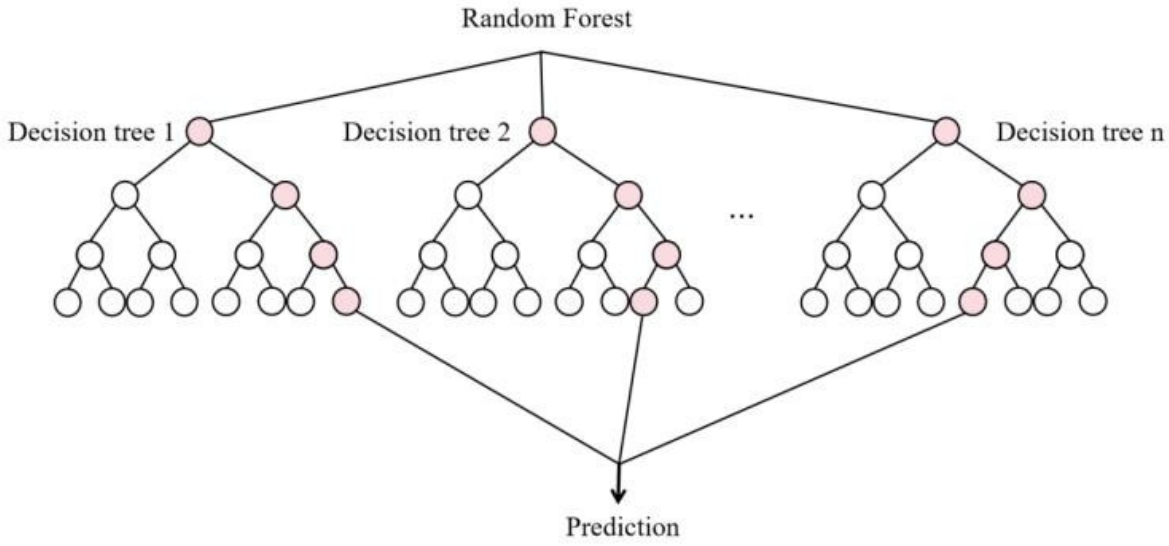


Figure 2.2 Structure of Random Forest

2.3.1.2 Extreme Gradient Boosting (XGB)

XGB uses multiple decision trees to show a distributed gradient boosting model. This model is well optimized within the gradient boosting framework. This is done to improve and achieve efficiency, flexibility and portability. Regularization terms are introduced to the objective function to reduce the variation between models; namely, it reduces issues with overfitting as models learn in a simpler manner. This model takes inspiration from Random Forest's approach wherein it supports column sub-sampling with faster computation as a result (Revathi et al., 2020).

2.3.1.3 Decision Tree (DT)

Decision Tree is the most common ML model as it is built on tree-based models that employ logic to predict an outcome. DT and its variations represent an alternate type of algorithms, each with individually parameterized algorithms (Liu et al., 2024). DT is a tree like structure with 3 main node types, which represents a process to contrast varying values on a data sheet of attributes, in turn determining the trend for the following decision step. State nodes represent values expected of an alternate solution, and by comparing the nodes an optimized result is found and selected. Through such algorithms attributes are divided and a DT's construction is completed through recursion. Additionally, during construction branching can be paused by pre- or post-running and prevent overfitting phenomena from occurring (Ahmad et al., 2018).

2.3.1.4 Gradient Boosting Regression (GBR)

Gradient Boosting Regression (GBR) is an enhancement learning algorithm based on DR, specifically made to solve regression problems. GBR has been shown to be particularly effective in altering prediction accuracy compared to only utilizing DT. The core fundamentals of GBR entail the initial training of a DT model on the dataset, followed by putting residual information within the training set. GBR trains following DT models through successive generations, merging them into existing models, following this it systematically adjusts the prediction results, while reducing errors from residual information. The final regression is the sum of multiple previous regression algorithms (Lu et al., 2018), as shown in the formula (2-1):

$$F_M(x) = \sum_{m=1}^{M} T(x, \theta_m) \tag{2.1}$$

The loss function for each weak classifier is defined as:

$$\hat{\theta}_m = \arg_{\theta_m} \min \sum_{i}^{N} L\left(y_i, F_{m-1}(x_i) + T(x_i, \theta_m)\right) \quad (2.2)$$

Where m represents the number of training iterations, x stands for the input data, and θ_m is the distribution weight vector. The model trains M times, with each iteration yielding a weak regression function T. $F_{m-1}(x_i)$ represents the current model.

2.3.2 Application of the machine model in water quality

2.3.2.1 Water quality prediction

Numerous models have been developed by using ML for the analysis of water quality and water security issues (Azrour et al., 2022). ML is expected to serve as a viable alternative to traditional water sampling, especially in measuring challenging-to-assess water quality parameters (Chowdhury et al., 2009; Shahi et al., 2020). In the field of water quality prediction, one of the most common research endeavors in ML is the prediction of Water Quality Index (WQI) (Aldhyani et al., 2020; Asadollah et al., 2021; Uddin et al., 2022). A newly integrated ML model, known as Extra Trees Regression (ETR), is employed for predicting the monthly WQI values in the Lin Village River in Hong Kong. Monthly water quality data, comprising chemical indicators such as biochemical oxygen demand and nitrite-nitrogen, along with physical indicators such as pH, turbidity, and temperature, are used as input features to construct the predictive model (Asadollah et al., 2021). Another study was explored to estimate the water quality index and water quality class (WQC) by ML, using four physical input parameters (Ahmed et al., 2019). ML also has a wide range of applications in predicting the chemical and physical eigenvalues of water bodies. The concentration of chlorophyll, DO, turbidity and conductivity were determined using an artificial neural network (ANN) algorithm with nonlinear autoregressive time series from a monitoring station in New York State (Khan & See, 2016).

In the Karouun River in southwest Iran, Mohammad Najafzadeh et al. (Najafzadeh & Ghaemi, 2019) used Multivariate Adaptive Regression Spline and Least Squares Support Vector Machine as water quality simulation methods to predict BOD_5 and COD. Utilizing various artificial neural network (ANN) models, the weekly concentration of nitrate nitrogen in the Sangamon River, situated close to Decatur, Illinois was forecasted. Also, the comparison showed that artificial neural network (ANN) models are better developed than linear regression in their study (Markus et al., 2003).

Additionally, utilizing sample data from 141 cases across small water distribution networks（SWDNs）and employing diverse ML methodologies, models were developed to predict three emerging disinfection byproducts (dichloroacetonitrile, chloropirrin, and trichloacetone) within SWDNs (Hu et al., 2023). Several ML models among other techniques, were constructed to predict 10 parameters related to irrigation water quality (IWQ) to assess the appropriateness of irrigation water (El Bilali & Taleb, 2020). Artificial neural network (ANN) has been utilized in an innovative way to predict water quality recovery, streamlining the resilience assessment process and obviating the need for parametric analyses traditionally employed in evaluating water quality recovery (Imani et al., 2021).

2.3.2.2 Classification

In the context of classification, ML can be categorized into supervised learning, semi-supervised learning, and unsupervised learning (The World Bank, 2021). When the training data includes corresponding labels, it is referred to as supervised learning, exemplified by algorithms such as Support Vector Machines and Random Forests. In cases where the training data lacks labels, it falls under the category of unsupervised learning,

which is used to explore the intrinsic structure of the data rather than predicting the specific output. If the training data consists of both labeled and unlabeled portions, it is termed as semi-supervised learning, with algorithms like self-training and co-training being relevant instances.

In the realm of water quality classification, extensive efforts have been dedicated to the application of ML. Dezfooli et al. employed three models, namely Probabilistic Neural Network (PNN), k-Nearest Neighbors, and Support Vector Machine (SVM), to classify the water quality levels based on the water quality parameters of 172 water samples from the Karun River in Iran (Dezfooli et al., 2018). Another study employed the SVM and Attribute Reduction (AR) algorithms to classify the water quality of the Mekong River. The input data consisted of monitoring data from 2008 to 2019, including turbidity, salinity, total coliform bacteria, biochemical oxygen demand, dissolved oxygen, ammonia nitrogen, and total nitrogen et al. Research in Bangladesh calculated the water quality index using the Weighted Arithmetic Index method based on data obtained from the Ghorashal Lake. Subsequently, a Gradient Boosting Classifier method was employed to categorize water quality into five classes ranging from "Excellent" to "Unsuitable for drinking" (Al-Razee et al., 2019). Another research used five ML classification methods, which were K nearest neighbors (K-Means), decision tree (DT), Naive Bayes, artificial neural network (ANN), and support vector machine (SVM), to predict WQC. The results showed that the decision tree and support vector machine classifier are the best prediction models with an error rate of 0% (Babbar & Babbar, 2017).

2.3.2.3 Heavy metals prediction

Neural Networks (NN) possess proficient data mapping capabilities, while Support Vector Machines (SVM) excel in effectively mapping small-sample datasets. These two ML models are commonly employed in current research for predicting heavy metal content.

ML for the prediction of heavy metals in soil has been studied in recent years. Utilizing the Random Forest (RF) model, the spatial distribution of soil-absorbable heavy metals in the arid regions of Iran for the years 1986, 1999, and 2010 was simulated. The results indicate that the RF model effectively predicts the distribution of heavy metals (Taghizadeh-Mehrjardi et al., 2021). In another study, the overall distribution of heavy metals in the soil in Hefei City, China, was predicted using RF, ANN, and SVM models. Soil characteristics, urbanization history, and the area of different land-use types were employed as predictive factors to estimate the concentrations of arsenic, zinc, lead, mercury, nickel, copper, chromium, and cadmium in the soil (Zhang et al., 2020).

ML methods can also be used to predict metals in the air. Research has been conducted applying meteorological factors and particulate matter (PM) concentration as predictive factors. Utilizing Multiple Linear Regression (MLR), Backpropagation Artificial Neural Networks (BP-ANN), and SVM, in conjunction with air PM data collected from Nanjing, China, rapid predictions of size-classified metals have been achieved (Wang et al., 2017). Additionally, based on four ML methods—MLR, BP-ANN, SVM, and RF, utilizing meteorological data, atmospheric pollutant data, and PM 2.5 data from the northeastern region of China for the years 2013 to 2018, predictive models for metal concentrations in atmospheric PM 2.5 were established (Lyu et al., 2023).

ML can also be used to predict the concentration of heavy metals in living organisms and sediment. One study used multiple linear regression models (MLR) and RF methods to estimate heavy metal concentrations in the muscle and liver tissues of psetta maxima maeotica which is a subspecies of turbot known as "a suitable biological indicator of heavy metal contamination in aquatic environments" (Petrea et al., 2020). A study in China used artificial neural network and support vector machine to predict heavy metal concentrations in sediments in Chaohu Lake, China, and analyzed its ecological risk index (Li et al., 2021).

However, the utilization of ML for predicting heavy metal concentrations in water bodies still lacks comprehensive research, primarily due to the limited availability of monitoring data. In Taihu region of China， Lu et al. used the physical and chemical indexes of surface water from drinking water sources, and combined ANN and SVM models to simulate dissolved substances, particulate matter and the concentration of heavy metals (Lu et al., 2019). Furthermore, in the southeastern part of Iran, research employed BPNN, generalized regression neural networks (GRNN), and MLR methods to predict the heavy metal concentrations (Cu, Fe, Mn, Zn) in the acid mine
drainage of the Sarcheshmeh porphyry copper deposit, while pH, Mg concentration and sulfate content served as input indicators (Rooki et al., 2011).

Currently, there is still a gap in predicting the concentrations of Fe and Mn in drinking water. Therefore, this study used four ML models: RF, XGB, DT and GBR to predict the iron and manganese content in drinking water in remote Indigenous areas in the province of BC, Canada. Meanwhile, this study used a variety of data augmentation

methods to address the limited water quality data problem which is the common challenge in terms of heavy metal modelling in drinking water.

2.3.3 Interpretability of machine learning

The interpretability of models is one of the most critical issues in ML applications. Interpretable approaches to the model can allow for explanations of how predictions are made. Which simply means, the purpose of interpretability is to turn the behavior of the model into understandable causal relations among various factors. Model-agnostic explanation systems offer a general framework for interpretability, enabling flexible selection based on the model itself, model features, and domain expertise. These model-agnostic tools enhance the credibility of ML applications in practice. Interpretable tools for models can be applied to any ML model after training. Model-agnostic methods, such as Accumulated Local Effects (Apley & Zhu, 2020), Local Interpretable Model-agnostic Explanations (Wang et al., 2021), and SHapley Additive exPlanations (SHAP) (Baptista et al., 2022), typically operate by analyzing feature input and output to provide insights into the model's behavior.

Currently, few studies have applied model interpretability to water quality prediction. Research employed RF model and pollution concentration data, including nitrate (NO_3-N), total phosphorus (TP), and Escherichia coli (E. coli), gathered from 1047 sampling stations in the Texas Gulf area to predict stream water quality under different levels of urban development scenarios. Model interpretation was conducted using the SHAP method to explore the influence of urban development patterns on stream water quality. The SHAP results highlighted the significance of indicators such as Landscape Division Index, Split Index, Maximum Patch Index and Patch Cohesion Index in shaping

stream water quality within the context of urban development. The study demonstrated that spatial variations in this pattern impact river water quality. The interpretability analysis of ML presented in this study suggests that the deterioration in river water quality can be attributed to the effects of urban rises (Wang et al., 2021).

2.4 Data augmentation (DA)

It is widely acknowledged that a substantial sample size is required in the practical application of artificial intelligence modeling. When constructing models with insufficient data, overfitting is prone to occur, resulting in decreased predictive accuracy and overfitting performance (Ma et al., 2023; Shen & Qian, 2022). Limitation hinders the ability to effectively interpret the variability of the target variable.

Data Augmentation (DA) represents a strategy for increasing the quantity of training samples (Connor et al., 2021; Iglesias et al., 2022), aiming to ameliorate challenges associated with insufficient samples (Shao et al., 2019) and imbalanced datasets (Zhao & Yuan, 2021) during the model training process. The generalization capability is enhanced through reducing overfitting and expanding the decision boundaries of the model (Fekri et al., 2019; Shorten & Khoshgoftaar, 2019). Through the DA process, the newly generated samples help to build a more robust and diverse training set, helping ML models learn a wider range of patterns, and improve generalization to unseen data.

DA methods are mainly divided into two categories: supervised and unsupervised. Supervised data augmentation methods include operations such as flipping, rotating, cropping, adding noise, SMOTE (Zhang et al., 2023), sample pairing (Inoue, 2018), mixing (Zhang et al., 2017) etc. Unsupervised data augmentation methods encompass Generative

Adversarial Networks (GAN) (Ma et al., 2023; Zhao & Yuan, 2021) and automatic data augmentation (Cubuk et al., 2018). The study involved multiple DA methods to enhance small sample datasets.

The first strategy introduced in this study is the integration of three conventional DA methods, namely Numerical Interpolation, Bootstrapping, and Noise Injection. The second method is Generative Adversarial Networks (GAN) which is a new technology based on the neural network (NN). These DA methods are introduced below.

2.4.1 Numerical Interpolation

Numerical interpolation is a commonly used technique utilized in data augmentation within ML. This strategy involves interpolating missing data or generating additional information from existing data. The process typically utilizes linear or polynomial interpretation as well as the nearest neighbor-based-interpolation in order to predict values based on the observed data points.

2.4.2 Bootstrapping

Bootstrapping is a method used to generate additional training samples by resampling the existing dataset.

When used for data augmentation, Bootstrapping involves selecting subsets from the original dataset causing replacement, in turn creating multiple bootstrap samples. Each sample is a variation of the original dataset, this introduces diversity to the training set. This process is very valuable as it allows generation of additional information without gathering new data.

2.4.3 Noise Injection

Noise injection is useful when there is lack of data or little diversity to the data. It is a regulation technique used to prevent a model from overfitting to data by exposing it to a broader set of input variations.

In this process, Gaussian noise or random jitters are added to the input data. The variation if caused can be applied to different types of data such as images, text or numerical data. With the introduction of controlled noise, the model becomes more resilient to small variations in the input. This makes it better at picking up on unseen or noisy data during training and inference.

2.4.4 GAN

Generative Adversarial Network (GAN) is an unsupervised learning algorithm first proposed in 2014 by Goodfellow et al. (Goodfellow et al., 2020). GAN is a class of artificial intelligence algorithms that consist of a generator and a discriminator of the confrontation game. They run the process simultaneously through adversarial training (Shao et al., 2019). The process involves the generator creating synthetic samples and the discriminator evaluating whether these samples are real or generated.

So far, data augmentation based on generative adversarial networks has predominantly been employed in the domains of image processing (Wang et al., 2019) and fault signal generation (Zhao & Yuan, 2021). It is worth noting that numerous research works have employed GAN in the medical field (Chen et al., 2020; Srivastav et al., 2021; Tyagi & Talbar, 2022). For example, utilizing GAN to augment image data for simulating pulmonary nodule shapes in chest X-ray which plays a significant prognostic role in early screening for lung cancer (Shen et al., 2023). Qin et al. developed a GAN-based screening

methods for melanoma and other skin diseases in dermo copy (Qin et al., 2020). However, there is limited research on augmenting small-sample continuous datasets.

In comparison to traditional data augmentation techniques, this method based on synthesis, although involving a more complicated process that typically requires training and learning, yields a more diverse set of synthetic samples.

2.5 Discussion and conclusion

Ensuring drinking water quality and security is a necessary step in the long-term to access clean drinking water which is a necessary basic human right. However, First Nations communities in Canada are still facing challenges in accessing clean drinking water due to the dispersed rural living patterns and specifically heavy metal pollutants in drinking water pose great risks to a person's physical health. It is important to build up a new method to detect metal concentrations in drinking water in rural First Nations communities to catch missing concentration data results from high-cost and time-consuming deficiencies from conventional lab analysis. ML is a potential method to simulate pollutants concentration in water, requiring large amounts of data to generate empirical models. Furthermore, this method can improve prediction accuracy efficiently by capturing the importance information of relevant features and is a better alternative than conventional methods in rural Indigenous areas; because they lack resources to utilize conventional methods. The performance of ML models is heavily affected by the quality and quantity of training data. Data augmentation is necessary to generate reliable synthetic data in a situation where there is a lack of water quality data. The applications of ML algorithms in predicting water quality and evaluating water security are broadly utilized, however the prediction of heavy

metals in drinking water still needs to be explored to fill up the blank of effective and reliable alternatives in terms of heavy metals detection in drinking water.

Chapter 3 Materials and Methods

3.1 Data processing and methods

3.1.1 Data sources and indicators

This study involved collecting aggregated groundwater data samples from five First Nations communities scattered across the province of British Columbia. The monitoring period spanned from 2019 to 2020, with water samples collected approximately at equal intervals. Subsequently, common water quality physical indicators and the concentrations of iron and manganese were analyzed for each sample, resulting in a total of 34 datasets.

The selected physical indicators, as detailed in Table 3.1—Total Dissolved Solids (TDS), conductivity, pH, turbidity, and hardness—were utilized as predictive factors in NI-BS-NI based data augmentation method, while in the GAN based data augmentation method, only pH, turbidity and hardness were used as input factors to accommodate some cases where limited input parameters can be detected caused by insufficient equipment in the practical scenarios. These chosen physical indicators possess the advantage of being detectable on-site using portable meters or simple titration, allowing for the rapid and cost-effective acquisition of concentration data.

To achieve a method that eliminates the complexity of laboratory testing processes and rapidly obtains metal concentrations, the concentrations of iron and manganese were selected in this study as the predicted indicators based on the practicality of the available dataset.

Table 3.1 Details of water quality physical parameters

Physical Parameter	Unit	Description	Importance	Detection Methods
TDS	mg/L	The total amount of dissolved solids in water, including inorganic salts, organic substances, and other dissolved materials.	Determines the level of dissolved pollutants in water, determining water suitability and usage.	Conductivity or evaporation methods
Conductivity	μS/cm	The ability of water to conduct electricity, primarily dependent on the dissolved ions present.	Monitors water salinity, pollution levels, and identify sources of water contamination.	Conductivity meters
pH	--	The measure of acidity or alkalinity in water.	Affects ecological balance and chemical processes, crucial for the survival of organisms and water safety.	pH electrodes
Turbidity	NTU	Indicates the amount and size of suspended particles in water.	Evaluates water clarity and visibility, detect potential water pollution.	Turbidity meters
Hardness	mg/L	Reflects the concentration of calcium and magnesium in water.	Affects water utility, plumbing systems, and equipment maintenance; crucial for sustainable water resource use.	EDTA titration or complexometric titration methods

3.1.2 Data preprocessing

Descriptive analysis of the raw data is beneficial for gaining an intuitive understanding of the predictor variables and the predicted target before modeling. In many instances, the quality of model prediction is directly linked to the raw data. For instance, the presence of outliers can significantly impact model results. Thus, conducting a descriptive analysis of the predicted target before modeling is a crucial step. In this study, the parameter analysis of the model primarily involves handling missing values and normalization of the data.

3.1.2.1. Missing value processing

The issue of missing data is one of the most common challenges during the process of data modelling. Common approaches to handling missing values include direct deletion, nearest neighbor imputation, linear regression fitting and so on. However, the effective handling of missing values within the model significantly influences the performance of the data model. Appropriately selecting suitable methods for dealing with missing values plays a crucial role in determining the overall effectiveness of the data model.

This study used the K-Nearest Neighbors (KNN) method to fill up missing values. The basic idea is to estimate the missing values based on the values of their nearest neighbors in the feature space (Zhang et al., 2017).

3.1.2.2 Normalization

Since the dataset contains variables with different ranges (i.e., the difference between maximum and minimum values), mean and standard deviation, data normalization is an important preprocessing step. Following the handling of missing values, the input

features underwent additional normalization using Equation 3.1. This normalization process aimed to standardize all input features, ensuring a consistent distribution of features.

$$X_i^* = \frac{x_i - \mu}{s} \tag{3.1}$$

In the equation, where x_i represents the value of input feature i, x_i^* is the normalized value of the initial x_i, μ is the mean of x_i, and s is the standard deviation of x_i.

3.1.3 NI-BS-NI based Data Augmentation

One of the challenges faced in this study is the limited data scale resulting from constrained water quality testing conditions in the SRR region. Therefore, data augmentation methods need to be set up to extract valuable information from the limited training data.

In the first strategy, we developed Numerical Interpolation-Bootstrapping-Noise Injection (NI-BS-NI) based data augmentation method to enhance the original data set. This method is a combination involves three traditional DA algorithms, which are Numerical Interpolation, Bootstrapping, and Noise Injection.

In the process of Numerical Interpolation, we interpolated for each feature to generate new 40 samples. Due to the limited size of the original dataset, which consists of only 34 samples, it is not conducive to generating too much data in this process.

Introducing the Bootstrapping DA method, 34 original samples were randomly selected, resulting in the generation of 66 data sets. In this process, new samples are generated while preserving the distribution of the data.

For data augmentation, Noise Injection is a commonly employed method. Noise can take various forms, such as Gaussian noise or uniform noise. Since the dataset primarily consists of continuous variables, Gaussian noise is often a preferable choice.

With a designated noise level of 0.05, 140 sets of new data were generated. Noise injection increases data diversity and helps to improve the generalization ability of the model.

After each round of data augmentation, the performance of the model was assessed using techniques such as cross-validation to determine whether data augmentation contributed to an improvement in model performance.

3.1.4 GAN based Data Augmentation

The second data augmentation method to generate new data in this study is Generative Adversarial Network (GAN). The overall framework is illustrated in Figure 3.1. It comprises four steps: dataset construction, GAN sample generation, removal of irrelevant values, and cross-validation loop iteration.

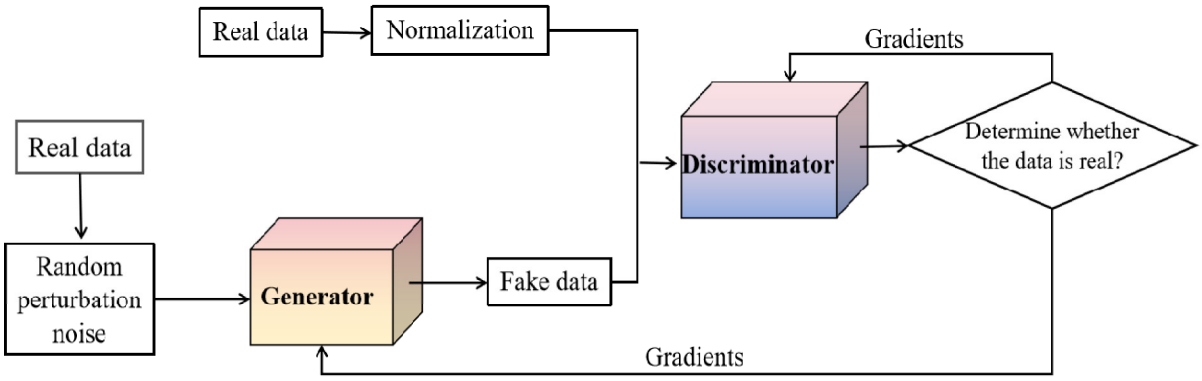


Figure 3.1 Structure of GAN

The specific process is outlined as follows:

Step 1：The original set of 34 data instances was divided into training and testing sets in an 80/20 ratio.

Step2：Constructing a GAN model for generating new samples involves two core network structures: the Generator and the Discriminator.

The Generator's objective is to generate synthetic data that closely approximates the real distribution, making it challenging for the Discriminator to distinguish the enhanced generated data. Meanwhile, the Discriminator's goal is to determine whether the data is real or fake, aiming to effectively distinguish between genuine and synthetic data. The Generator and Discriminator engage in an adversarial process, iteratively enhancing their respective discrimination or generation capabilities. When the loss functions for both the generated and discriminative networks converge, the Discriminator becomes reasonably adept at authenticating real samples typically. However, certain generated data may still be misclassified as real, which means the Generative has learned the properties of real samples and can produce plausible synthetic data. The learning rate is 0.001. The batch size is 128 in this process. The optimization during the training of the Generator and Discriminator utilizes the Adam algorithm.

Step 3: Delete the similar samples and unreasonable samples generated in step 2. The similar sample removal procedure is based on the Euclidean distance compared with the set threshold. The threshold was set at 0.5.

Step 4: A 3-fold cross-validation approach was applied at each iteration of the GAN, as shown in Figure 3.2, to assess the quality of the generated samples. MAE_{GAN} (Mean Absolute Error from GAN-generated data) was computed through three-fold cross-validation, while the MAE_{train} was obtained through three-fold cross-validation on the training set data from step 1. Subsequently, a comparison between the two was conducted. If MAE_{GAN} is smaller than MAE_{train}, indicating that the newly generated data exhibits a smaller mean absolute error in the three-fold cross-validation process compared to real data,

then the generated samples are of higher quality. The iteration concludes when the iteration count is greater than or equal to 10, resulting in the final set of data.

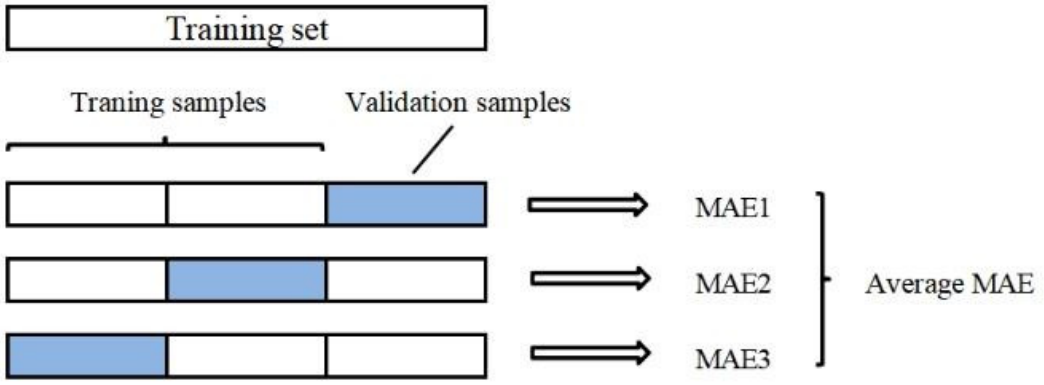


Figure 3.2 The 3-fold cross validation diagram

By learning the distribution of real samples to generate synthetic data, during each training iteration, a set of random noise z ~ N ~ (0,1) is input into the generator to produce fake samples. The discriminator assesses the authenticity of the samples and assigns scores. The generator's objective is to deceive the discriminator into classifying fake samples as real, while the discriminator aims to distinguish between fake and real samples. Through adversarial training, the goal is to make the distribution of generated samples approach that of real samples. The objective function of GAN is expressed by the following formula:

$$\min_{G}\max_{D} V(D,G)=E_{x\sim P_{data}}\left[\log D(x)\right]+E_{z\sim P_{g}}\left[\log(1-D(G(z)))\right] \tag{3.2}$$

In the expressions: P_{data} and P_g represent the distributions of real samples x and random noise z, respectively. G(z) denotes the generated pseudo samples, D(x) signifies the probability of real samples being judged as authentic, and D(G(x)) indicates the probability of pseudo samples being judged as authentic.

The GAN model incorporates two distinct loss functions designated for training the generator and discriminator network. The loss functions of the generator and the discriminator are expressed as the mean absolute error (MAE) and the binary cross-entropy (BCE). Their definitions are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.3}$$

$$\mathrm{BCE} = -\frac{1}{n}\sum_{i=1}^{n} [y_i \cdot \log p(y_i) + (1 - y_i) \cdot \log(1 - p(y_i))] \tag{3.4}$$

where n represents the sample size, $\hat{y}_i$ denotes the predicted values, y_i represents the actual values, and $\bar{y}_i$ represents the mean of the actual values.

3.2 Machine Learning Modelling

3.2.1 Model development

In this study, Python version 3.9 is utilized, and the ML model library is retrieved from Scikit-learn. Python is a high-level scripting language that combines interpretability, compatibility, interactivity, and object-oriented programming. It is also the most popular language for ML, featuring the most comprehensive and up-to-date ML frameworks such as TensorFlow, PyTorch, and others.

In this study, Random Forest (RF), Gradient Boosting Regression (GBR), Extreme Gradient Boosting (XGB), and Decision Tree (DT) were selected for modeling by study requirements and model characteristics. These 4 integrated tree models have good results interpretative, and separately belong to different integration methods.

After preprocessing and augmenting the dataset, the data points are further divided into training and test sets with the same distribution for multiple training sessions. In this study, the dataset is divided into a training set and a test set with an 8:2 ratio.

Simultaneously, a 5-fold cross-validation method is employed during the model training process for validation to avoid overfitting and improve the predictive power of the applied ML methods.

Single-parameter optimization and multi-parameter optimization were both utilized in modelling to find the best optimization method in this study. After combing with three traditional data augmentation processes, multi-parameter optimization was used to predict iron and manganese, while single-parameter optimization was studied after GAN data augmentation in ML modelling.

3.2.2 Model hyperparameter optimization

Hyperparameter tuning is very important to obtain the best model. The best set of hyperparameters plays an important role in model reliability and adaptability. The most common methods for hyperparameter tuning include manual search, grid search, and Bayesian optimization. In this study, the grid search method is employed for hyperparameter tuning, as its precision when dealing with a limited number of feature inputs particularly. During the parameter tuning process, model parameter optimization is conducted using the average results of the five subsets from each validation iteration in the five-fold cross-validation.

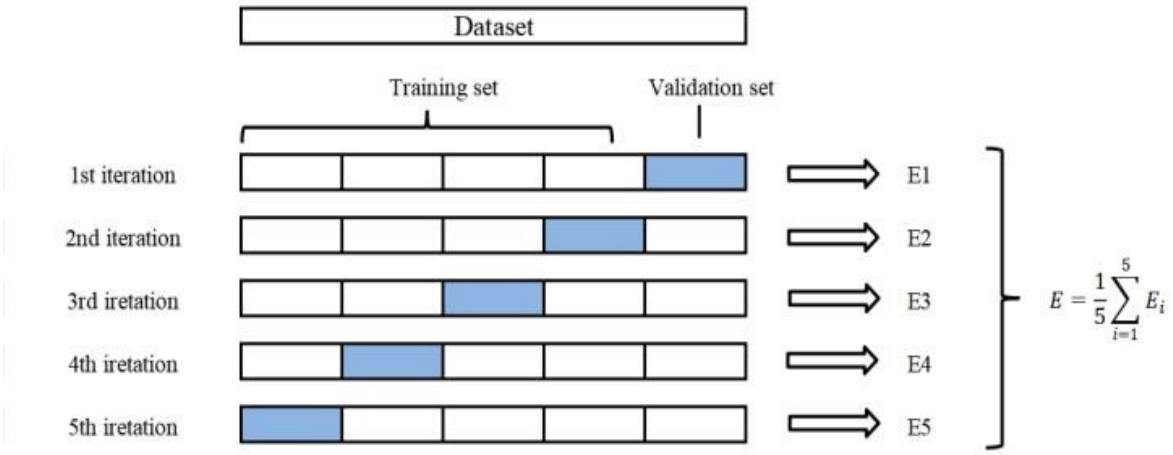


Figure 3.3 Diagram of five-fold cross-validation

3.2.3 Model evaluation

For a ML predictive model, classification predictions and numerical predictions have different discriminant formulas. The discriminant formulas for predictions of the same type generally exhibit a fundamental consistency, which is also a manifestation of the model's robustness. Common metrics for evaluating the accuracy of each model include the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

RMSE represents the errors between predicted samples and actual samples. MAE measures the average absolute difference between the predicted values and the actual values in a dataset. Lower RMSE values and MAE values indicate more accurate predictions of the model, with 0 being the optimal score (perfect predictions).

The calculation formulas for RMSE and MAE are shown in Formula 3.5 and 3.6:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{3.5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.6}$$

The study also calculated the coefficient of determination (R^2) between the predicted values and observed values of the test set to assess the quality of the predictions. R^2 acknowledged the fitting ability of established models. A higher R^2 value indicates a better fit of the model.

The formula of the R^2 is shown in Equation 3.7:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} \tag{3.7}$$

In the above equations, n represents the sample size, $\widehat{y}_i$ denotes the predicted values, y_i represents the actual values, and $\overline{y}_i$ represents the mean of the actual values.

3.3 Model interpretability analysis

This study attempts to use ML interpretability methods to explain and elucidate the model's prediction results. Following the more reliable predictions for iron and manganese, to further investigate the impact of predictor variables on water quality, it is necessary to explore the magnitude, direction (positive or negative), and interactive effects of predictor variables on the concentrations of iron and manganese in drinking water. This is beneficial for precisely identifying the primary factors contributing to variations in target water quality parameters in specific contexts. ML model interpretability methods can effectively explain tree models.

3.3.1 Impact magnitude of predictors

Feature importance is often the first step in interpreting models in data mining. ML's feature importance reflects the magnitude of the impact of features on model predictions (Ibrahim et al., 2019). Calling “the feature importance” in the Scikit-learn

library allows direct retrieval of feature importance. It ranks features based on their frequency of use within the classifier and outputs a graph illustrating their ranking in importance. Feature importance is very intuitive thus making it very easy to understand the weightings of each feature and recognize their importance on model predictions. However, feature importance is model-dependent as such, models that generate rankings differently will lead to differing results. Additionally, feature importance does not capture feature interactions following that rankings may be influenced by noise within data, ultimately resulting in information that deviates from the original situation.

3.3.2 Interactive effects of predictors

This study used Partial Dependence Plot (PDP) to investigate the interactions between indicators. Other common methods used for revealing the interaction effects of predictive factors in ML models include Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015) and Local Interpretable Model-agnostic Explanations (LIME) plots (Ribeiro et al., 2016). PDP illustrates the marginal effects between one or two features and the prediction outcomes of the ML model, showing whether the relationship is linear, monotonic, or more complex (Nie & Wager, 2021; Yan et al., 2020). Unlike feature importance in header 2.3.1, which indicates the numerical magnitude of a feature's impact on the model, PDP presents the relationship between features and the impact on prediction outcomes. This paper considers the impact of individual factors on prediction outcomes and the combined effects of two factors, specifically examining the synergistic effects of two features on predictions.

Partial Dependence Plots are intuitively defined, easy to understand, highly interpretable, and computationally efficient. They can effectively explore the joint impact

of two features on the model's predictions. However, their drawback is that they can describe the impact of only two features simultaneously. Moreover, when there is a strong correlation between two features, the results may exhibit bias.

3.3.3 Impact direction of predictors

When we understand that a certain predictor has a significant impact on water quality parameters, it is even more important to know the direction of this impact, in other words, as this predictor increases, do the water quality parameters increase or decrease.In this context, we utilize SHapley Additive exPlanations (SHAP) plots to observe the positive and negative impacts as well as the magnitudes. SHAP values are based on cooperative game theory and Shapley values (Winter, 2002), assigning a value to each feature based on its marginal contribution to different possible feature combinations. Additionally, it provides the contribution of each feature data point to the predicted value (Lundberg & Lee, 2017). Thus, it can be applied to individual predictions.

Feature importance as mentioned in 3.3.1 provides a high-level overview of the relevance of features across the dataset. SHAP values offer a detailed breakdown of how each feature contributes to a specific prediction, considering interactions among features. Both can be valuable tools for model interpretation and gaining insights into the factors driving model predictions.

Chapter 4 Results and Discussion

4.1 NI-BS-NI Based Modelling

4.1.1 Statistical analysis

4.1.1.1 Statistical summary

Statistical analysis plays a foundational role in subsequent model establishment and interpretable analysis. Statistical information of the data can intuitively demonstrate the data distribution, dispersion, and central tendency. Based on the data characteristics and information, adjustments can be made to the ML modelling operation.

In this study, statistical analysis is performed on both the raw data and the augmented data to examine the reliability of data augmentation methods. This involves assessing whether the data augmentation is based on the characteristic distribution properties of the raw data, rather than generating data unrelated to the original dataset and leading to untrustworthy predictive outcomes.

Tables 4.1 and 4.2 respectively present the statistical information of the five predictive factors and the predicted indicators of raw data and the data after augmentation through three traditional methods. The data has increased from 34 sets to 276 sets. The data covers a wide range, allowing for robust model results with strong generalization capabilities.

Table 4.1 Statistical summary of raw data

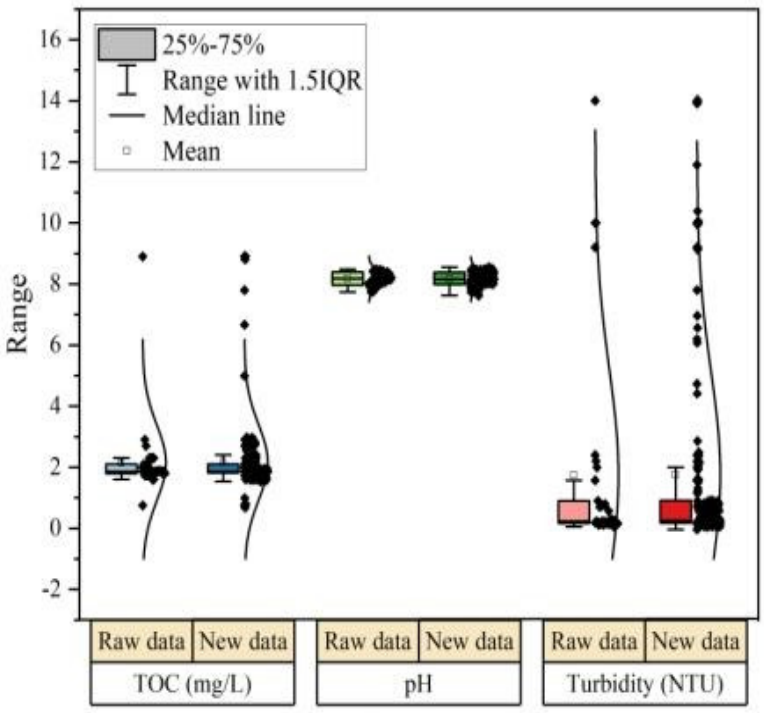
	TDS (mg/L)	pH	Conductivity (µS/cm)	Turbidity (NTU)	Hardness (mg/L)	Fe (mg/L)	Mn (mg/L)
count	20	34	20	34	34	34	34
mean	325.050	8.164	535.300	1.735	174.521	0.187	0.195
std	196.970	0.232	315.613	3.477	120.023	0.411	0.342
min	89.000	7.740	163.000	0.050	0.510	0.003	0.0001
25%	240.750	7.968	369.750	0.173	82.625	0.011	0.003
50%	299.500	8.175	483.500	0.235	155.000	0.036	0.030
75%	345.000	8.385	567.750	0.875	260.500	0.158	0.203
max	870.00	8.480	1400.000	14.000	470.000	2.170	1.160

Table 4.2 Statistical summary of NI-BS-NI augmented data

	TDS (mg/L)	pH	Conductivity (µS/cm)	Turbidity (NTU)	Hardness (mg/L)	Fe (mg/L)	Mn (mg/L)
count	276	276	276	276	276	276	276
mean	287.282	8.158	480.981	1.646	176.618	0.182	0.193
std	161.066	0.229	255.833	3.223	114.049	0.382	0.317
min	88.915	7.587	162.963	0.0001	0.510	0.0001	0.0001
25%	146.089	7.960	265.356	0.172	93.516	0.011	0.003
50%	260.000	8.170	470.000	0.240	157.015	0.062	0.041
75%	348.585	8.379	570.000	0.908	258.979	0.158	0.245
max	870.052	8.499	1400.038	14.088	470.078	2.262	1.232

It can be observed that after data augmentation, the trend statistics (mean and percentiles), the measures of dispersion (standard deviation), and the distribution statistics (maximum and minimum values) are relatively consistent compared to raw data, falling within a reasonable range. This proves the effectiveness of using this method for data augmentation.

Figure 4.1 shows the comparison boxplot of the raw data and the new data generated after the NI-BS-NI data augmentation. Figures show the distribution of the data, including the median, the upper and lower quartiles, and other information intuitively. The results showed that the data augmentation operation had no obvious effect on the data distribution, and 25% -75% of the data of each factor are still concentrated, and no obvious outliers appeared.



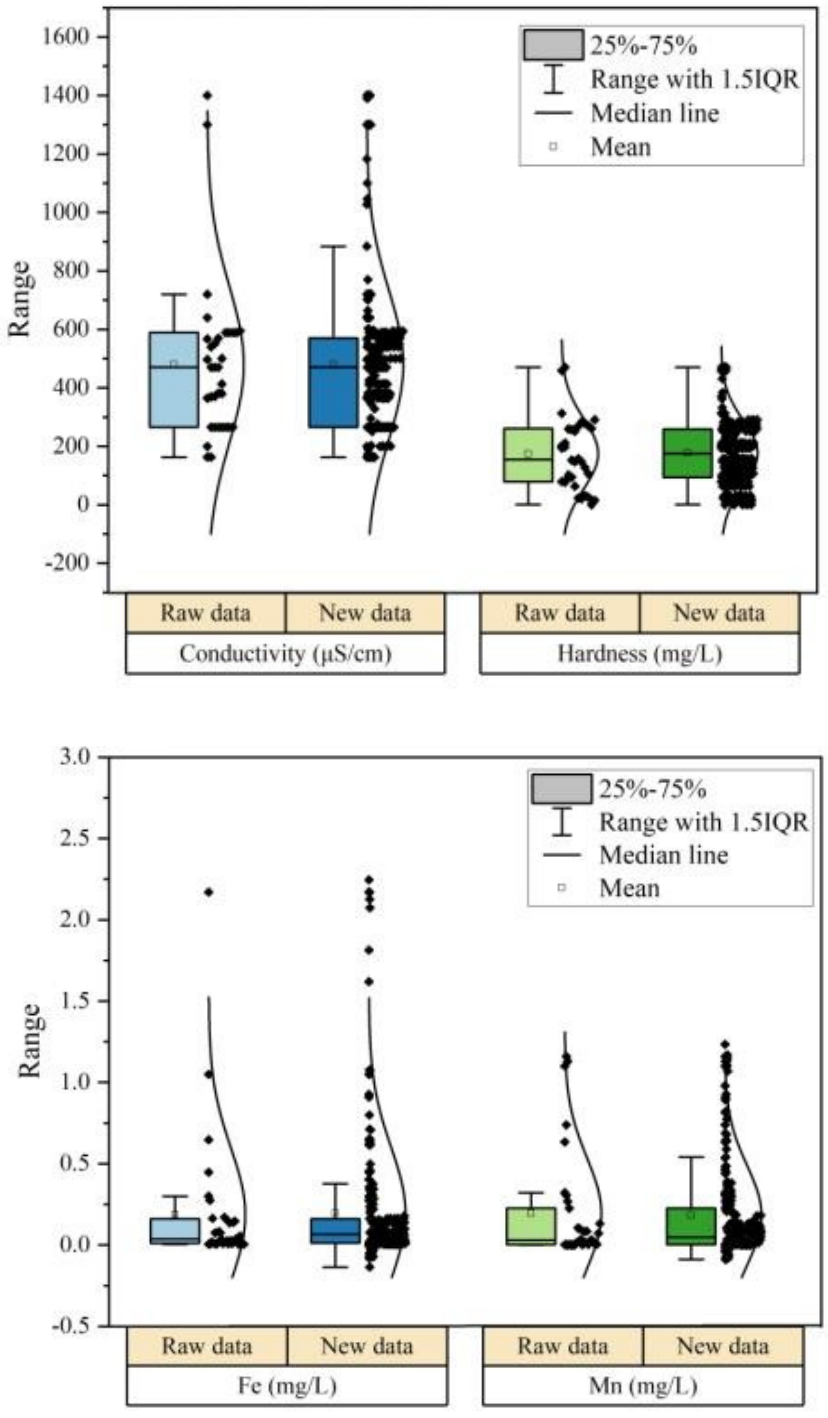


Figure 4.1 Comparison boxplots of raw data and NI-BS-NI augmented Data

4.1.1.2 Correlation analysis

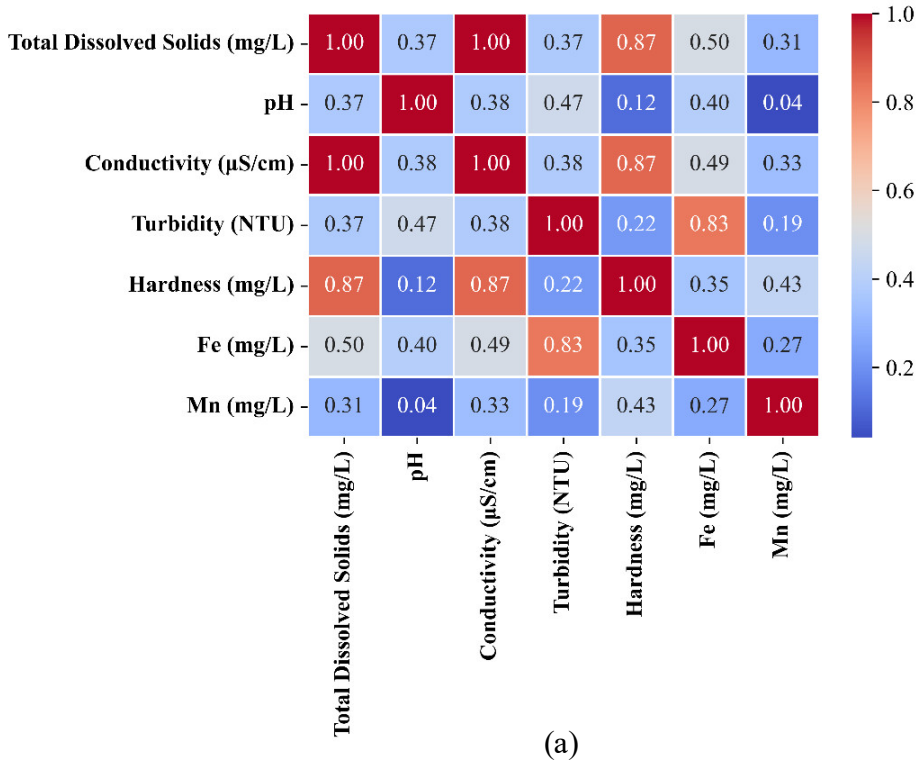
For the original and augmented data, a correlation analysis of the data was conducted to detect the correlation between numerical independent variables. Pearson correlation coefficient heatmaps were generated, as shown in Figure 4.2.

As we can see, the correlation coefficient between Hardness and Conductivity is 0.87 (raw data). The strong correlation between them may be attributed to the influence of dissolved substances in water. Dissolved minerals such as carbonates, sulfates, and chlorides in water enhance the hardness in water, meanwhile, the water's conductivity also increases because calcium and magnesium irons commonly exist in hard water. Therefore, hardness and conductivity exhibit a positive correlation. Similarly, hardness shows strong correlation with TDS may be caused by calcium and magnesium irons can also react with other substances in the water to form a precipitate or suspended substances to increase the TDS.

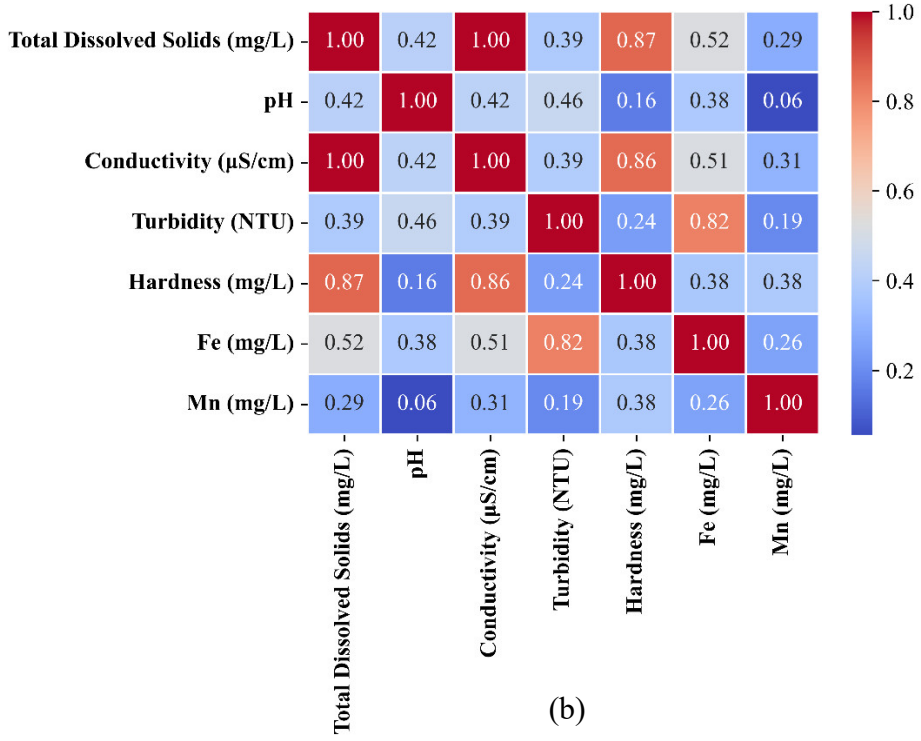
In addition, there is a strong correlation between the concentration of iron and turbidity. The correlation coefficient is greater than 0.8.

The reason for the strong correlation between iron concentration and turbidity in water may be because iron in the form of suspended particles in water increases the turbidity, which is the measure of suspended particles.

The correlation between other variables is not significant strong, which means that most variables contain unique information, and there is no excessive information overlap.



(a)



(b)

Figure 4.2 Pearson Correlation Heatmap of (a)raw data (b) NI-BS-NI augmented Data

4.1.2 Model results

Table 4.3 Performance of NI-BS-NI based models and best hyperparameters

Model	Output Variable	Best Parameters	Train R^2	Test R^2	Train RMSE	Test RMSE
RF	**Fe**	{'max_depth': 40, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 10}	0.992	0.969	0.035	0.040
RF	**Mn**	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 20}	0.968	0.943	0.055	0.093
XGB	**Fe**	{'learning_rate': 0.1, 'max_depth': 20, 'n_estimators': 100, 'subsample': 0.7}	1	0.974	0.002	0.036
XGB	**Mn**	{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 100, 'subsample': 0.7}	0.9999	0.933	0.002	0.101
GBR	**Fe**	{'learning_rate': 0.1, 'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 50}	0.999	0.969	0.01	0.04
GBR	**Mn**	{'learning_rate': 0.1, 'max_depth': 20, 'min_samples_leaf': 4,	0.9999	0.979	0.002	0.056

		'min_samples_split': 5, 'n_estimators': 100}				
DT	**Fe**	{'max_depth': 50, 'min_samples_leaf': 2, 'min_samples_split': 2}	0.997	0.95	0.022	0.051
DT	**Mn**	{'max_depth': 50, 'min_samples_leaf': 2, 'min_samples_split': 2}	0.992	0.976	0.027	0.06

After augmenting the raw data by using NI-BS-NI method, this study used four ML algorithms, which includes Random Forest (RF), Extreme Gradient Boosting (XGB), Decision Tree (DT) and Gradient Boosting Regression (GBR) to predict iron and manganese concentration in drinking water and find the best parameters during modelling to obtain the best performances of each model.

The results of the modelling achieved good results as shown in Table 4.3. Compared with the other three models, GBR has the best performance in all the simulation results. In row GBR the Train R^2 of iron and manganese are just neath of 1, and Test R^2 achieved ideal results, which are 0.969 and 0.979 respectively. And RMSE are all lower than 0.01, and as such, the errors between predicted data and actual data are minimal.

Although XGB has the highest Train R^2 result, the Test R^2 is comparatively low compared to the training results, when predicting manganese XGB scores especially low in R^2 (0.933), which may cause by overfitting phenomenon. Besides, the Test RMSE of XGB model didn't show a good result in terms of predicting Mn concentration ($RMSE_{Mn}$=0.101). In addition, the performance of RF and DT models showed

comparatively poor simulation performance, reflected in lower R^2 and higher RMSE scores. A possible reason may stem from the fact that RF and DT models are primitive compared to GBR model; as such, they perform poorly when simulating data with non-linear and complex characteristic relationships.

In addition, scatter diagrams and regression lines of four models in predicting iron and manganese were also produced to visualize the performances of simulation, as shown in Figures 4.3-4.6. These diagrams show the consistency of predicted values and actual values of metals concentration. According to the given reference line in the diagrams, all 4 models present great predicative performances, which are expressed in scatter points being close to reference lines in each diagram. The GBR model showed excellent performance in predicting Fe and Mn, as shown in Figure 4.5, most scattered points fell on the reference line.

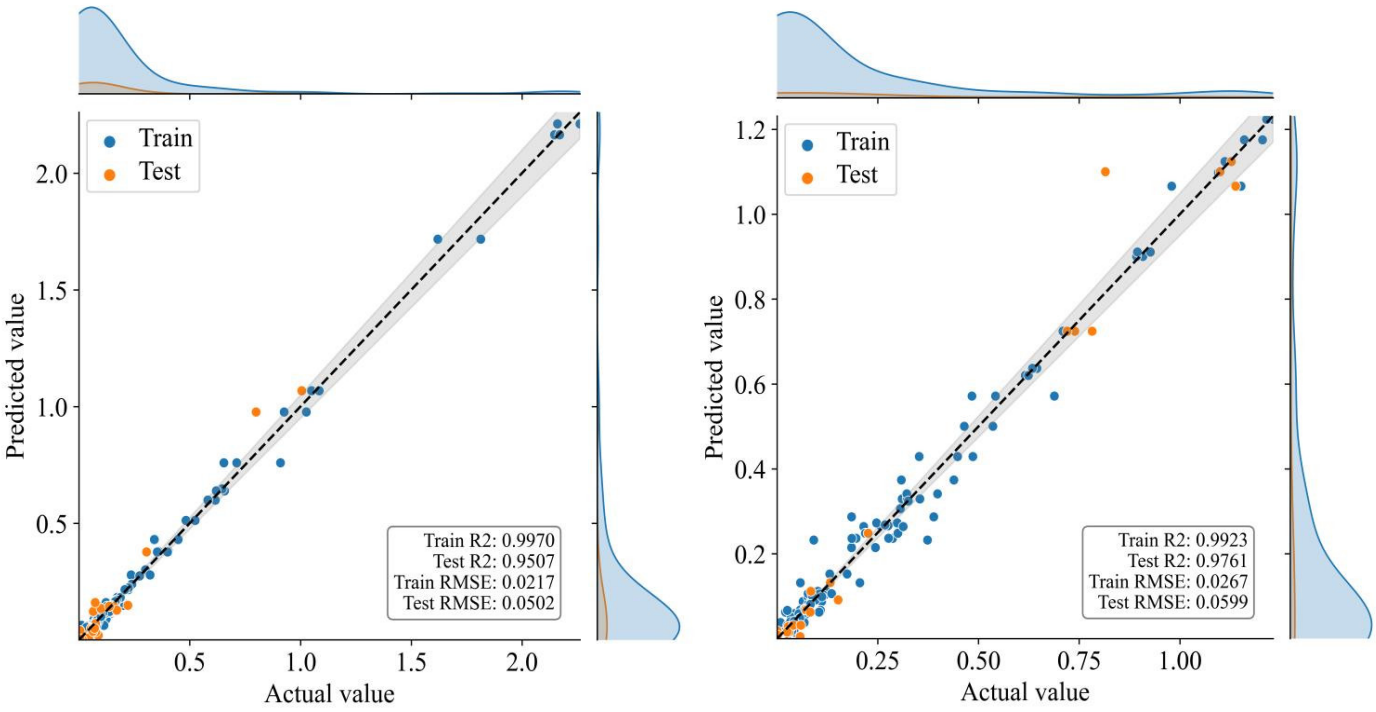


Figure 4.3 Scatter regression plots of (a)Fe (b)Mn of DT

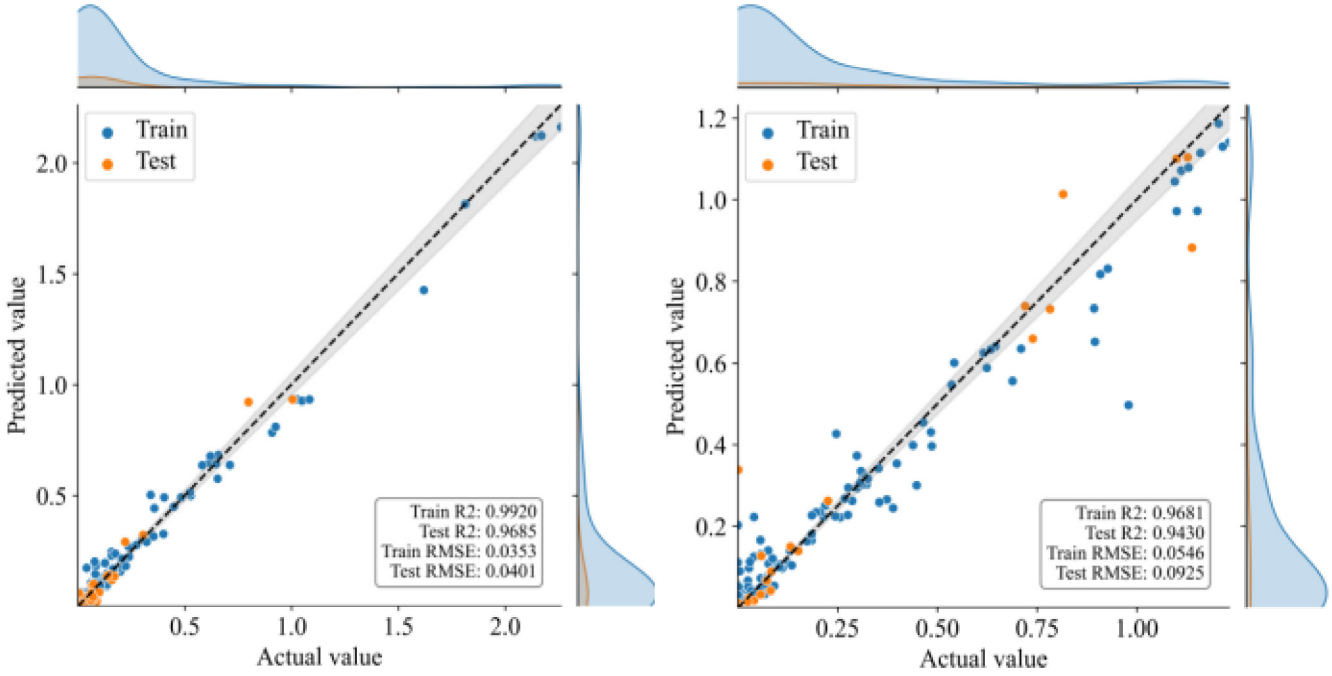


Figure 4.4 Scatter regression plots of (a)Fe (b)Mn of RF

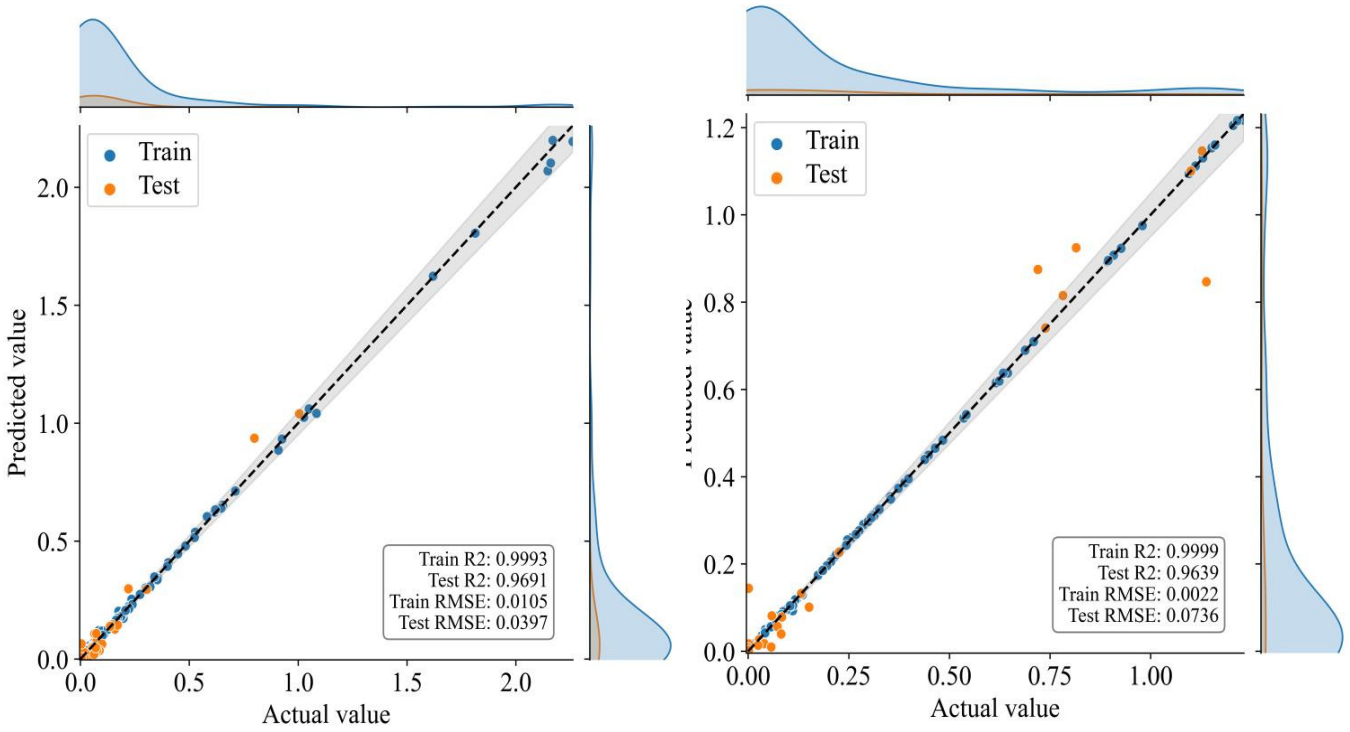


Figure 4.5 Scatter regression plots of (a)Fe (b)Mn of GBR

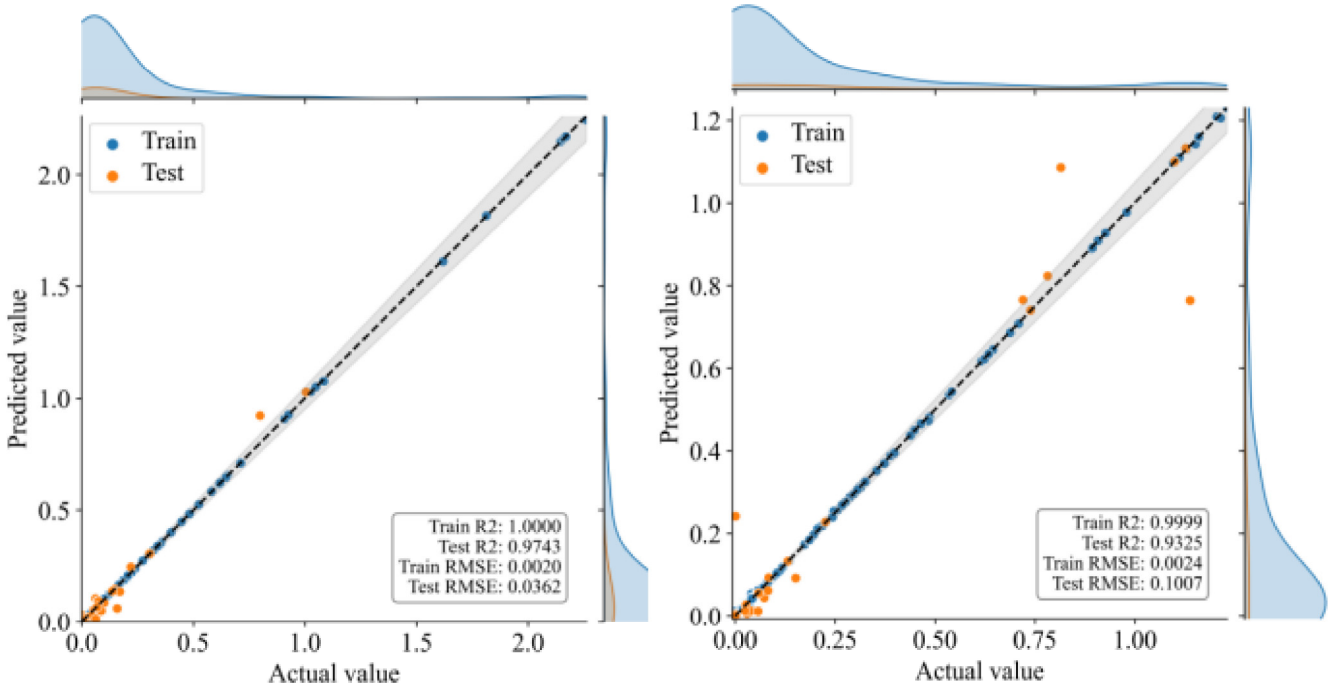


Figure 4.6 Scatter regression plots of (a)Fe (b)Mn of XGB

4.1.3 Interpretable analysis

4.1.3.1 Feature importance

In this section, the best model GBR in predicting iron and manganese was selected to present the effect of the 5 predictors on Fe and Mn concentration. Obtaining feature importance by directly calling the 'feature_importance' in the model to output Feature Importance Ranking Plots, as shown below.

Figure 4.7 shows the ranking of 5 input predictors feature importance in predicting Fe and Mn, also can be explained as the ranking of the influence size of each feature. From top to bottom, the importance of each features decreases gradually. For the Fe prediction, turbidity has an outstanding impact, with the importance scores over 0.8. Followed by hardness, TDS, conductivity and pH, with low importance. Likewise, to Mn prediction,

conductivity is the most important feature, followed by hardness and turbidity, also show some impact to Mn prediction. TDS and pH still show little importance as they do to iron.

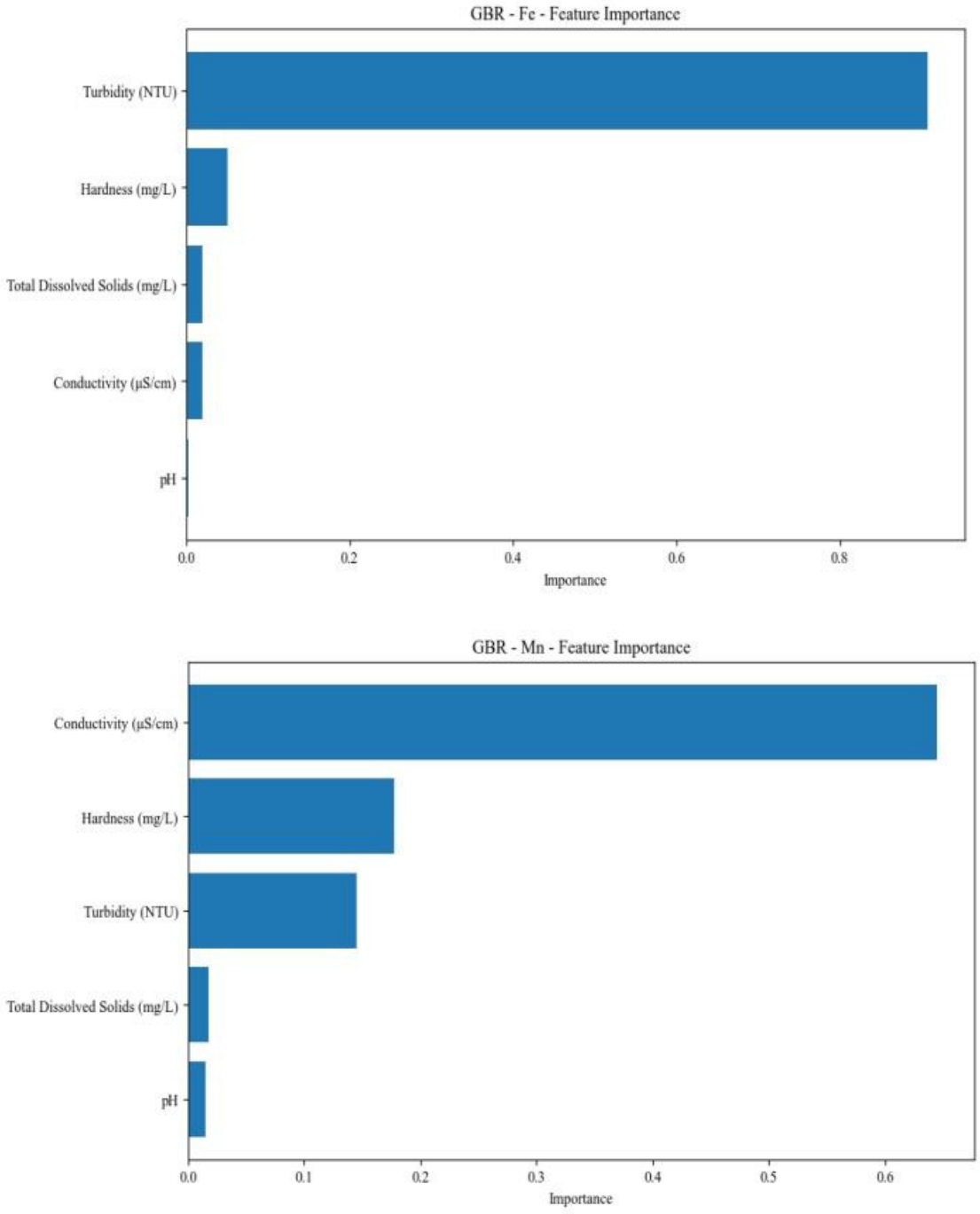


Figure 4.7 Feature importance of GBR model after NI-BS-NI augmentation

4.1.3.2 Partial dependence

In this section, the dependence of each feature on the metal's prediction will be explained. The explanation of how each feature affects the prediction distinguishes partial dependence from feature importance discussed in section 3.3.1.

Take XGB model in this simulation as example, the 5 partial dependence plots were created to explain the effect of the 5 input features on iron and manganese concentration. In Figure 4.8, the x-axis represents the range of variation of each input variable, the y-axis means the predicted values.

The partial dependence plot proves the result we obtained from correlation analysis: iron and turbidity show great correlation and have little correlation with other features. As shown in Figure 4.8, the predicted iron concentration increases significantly with the increase of turbidity. For TDS, pH, conductivity and hardness, the iron concentration stays in the horizontal state basically, indicating that these four physical characteristics have very limited impact on the iron concentration.

From the dependence plot of manganese, the partial dependence of all 5 features shows greater effect compared to iron partial dependence: the Mn values fluctuates in every plot. Mn concentration decreases from 0.3 to 0.1 mg/L with the TDS increases until 400 mg/L, then stabilizes at around 0.1. The concentration of pH and manganese show a small negative correlation. As for conductivity, it shows a significant positive correlation, the Mn concentration increases from 0.1 to 0.4 mg/L rapidly when the conductivity changes from 300 to 600 μS/cm. The Mn concentration generally increases with the turbidity increases as shown in the figure. When hardness is around 250 mg/L, the concentration of Mn peaks at about 0.35mg/L.

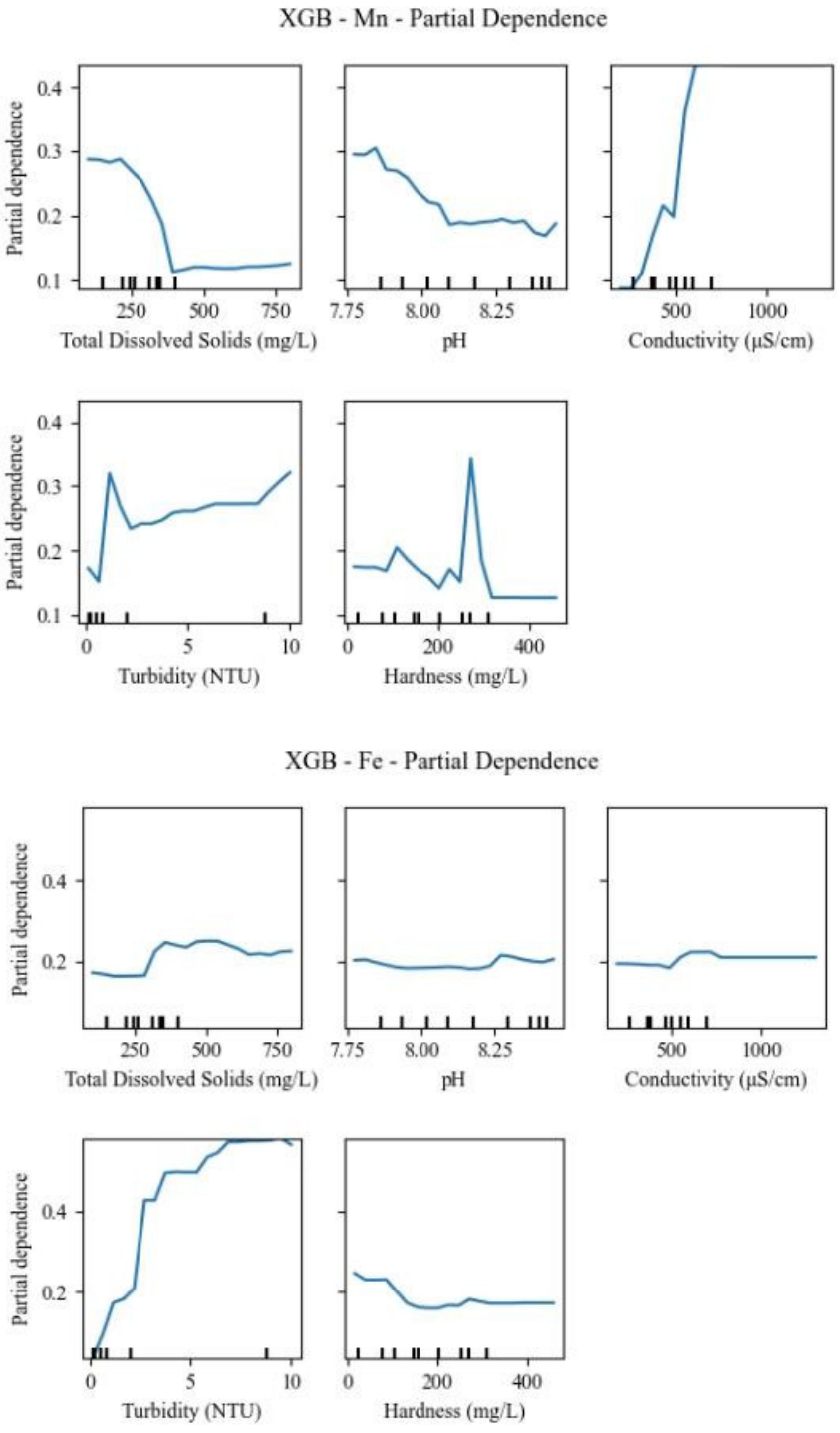


Figure 4.8 Partial Dependence Plots of XGB model

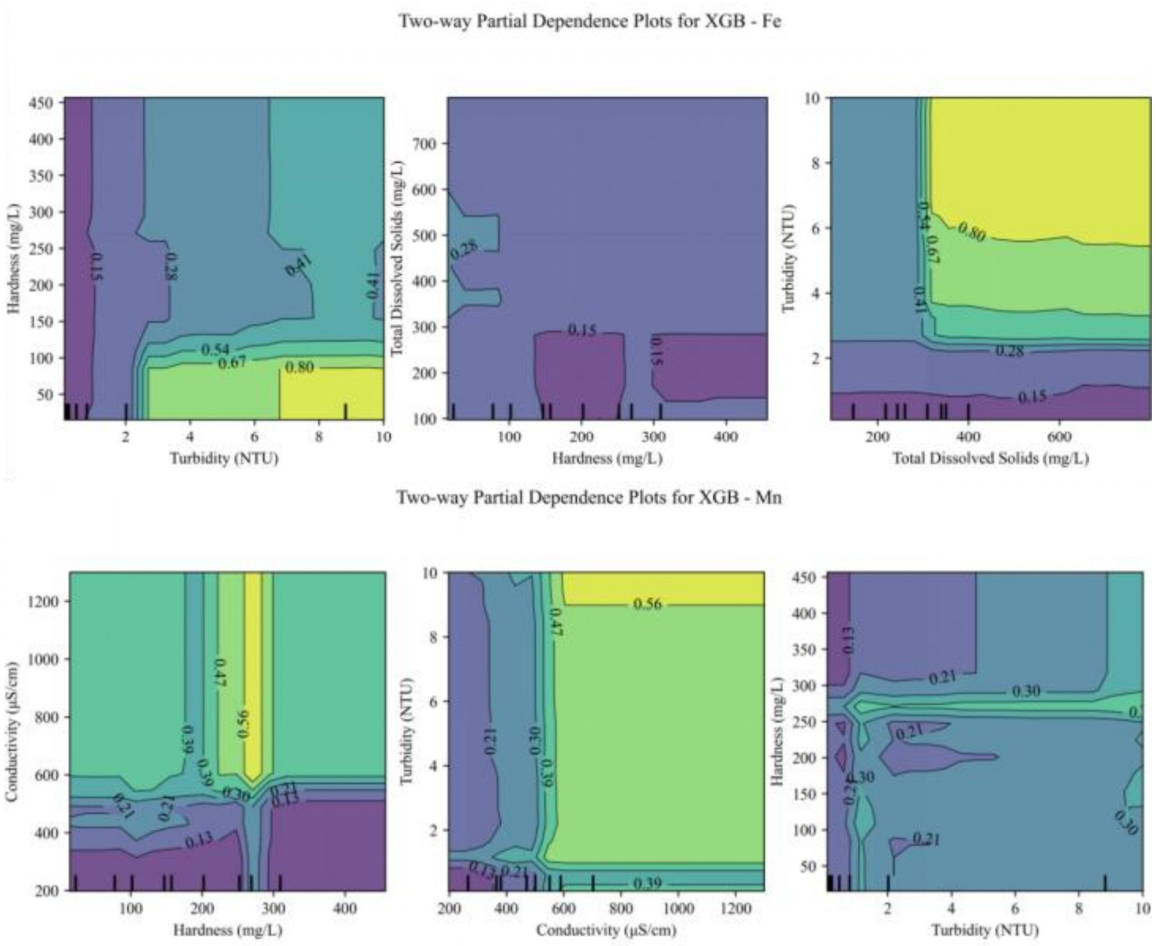


Figure 4.9 Two-way Partial Dependence Plots of XGB model

After analyzing the individual features partial dependence, the synergistic effect of two factors on the prediction of iron and manganese also need to be considered. According to the correlation analysis in section 4.1.1.2, most of the indicators show generally low correlations. Thus, it is significant to explore the combined effect of two variables to predictive values. Figure 4.9 shows two-way partial dependence of GBR model to iron and manganese prediction. By the horizontal and vertical axes represent the variation range of the two features, and the third dimension is represented by the color differences: the yellow

color block shows the greater predicted concentration, and purple indicates a small concentration.

By the common influence through two features, the greater the turbidity and the less the hardness, causing the greater the concentration of iron. The combined effect of TDS and hardness had little effect on iron concentration. When the turbidity is greater than 6 NTU and the TDS is greater than about 320 mg/L, the iron concentration will increase to 0.8 mg/L or above. For manganese, the concentration is maximum with increasing conductivity and a hardness between 250-280 mg/L. The synergy between the turbidity and the conductivity plot indicates that the greater their concentration, the greater the concentration of manganese. However, the combined effect of hardness and turbidity had little effect on the manganese concentration.

4.1.3.3 SHAP analysis

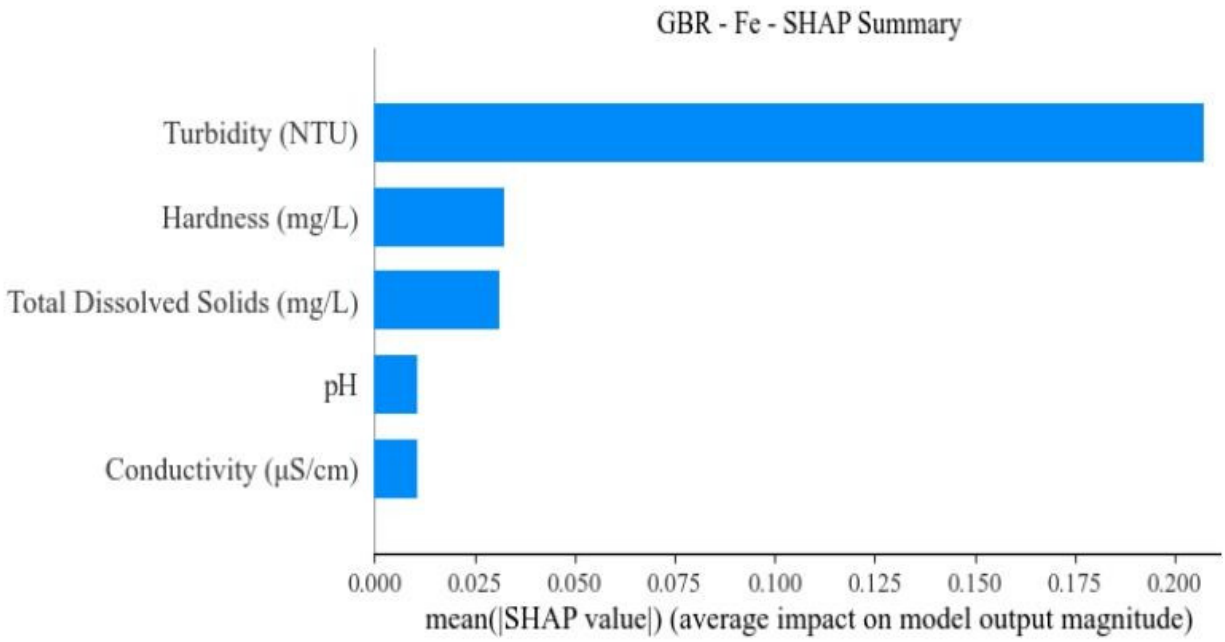


Figure 4.10 SHAP summary plot of Fe of GBR model

In this section, the importance of each feature will be expressed by SHAP value. Figure 4.10 indicates the ranking of average impact on model output magnitude determined by the mean absolute value of the SHAP (|SHAP value|). In the five input features, the effect of turbidity is most pronounced for iron prediction, next by hardness and TDS, with an equal magnitude of impact. The predicted effect of pH and conductivity on iron is not obvious.

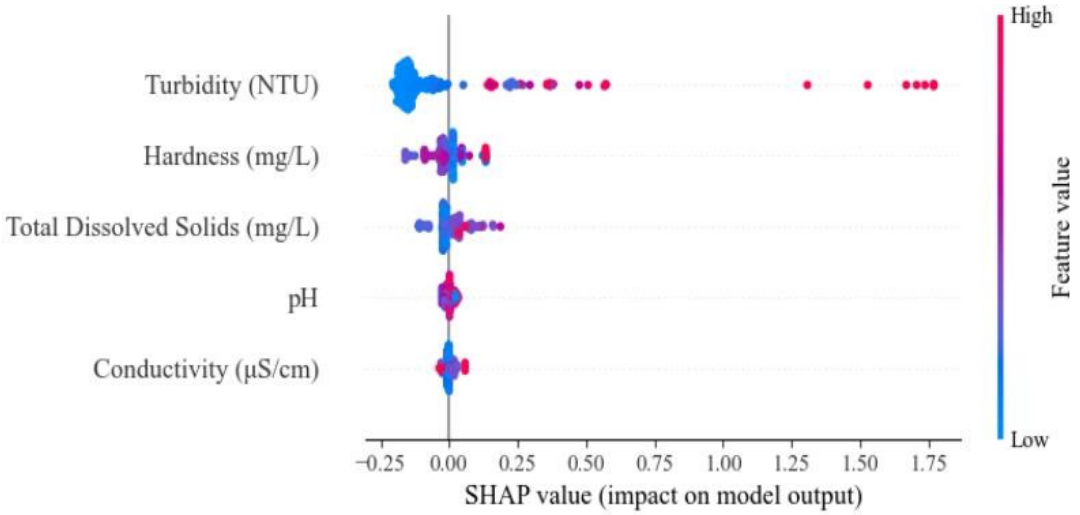


Figure 4.11 SHAP scatter diagram of Fe of GBR model

Detailed explanation will be analyzed by SHAP scatter diagram as shown in Figure 4.11. The horizontal axis represents the different SHAP values, while positive values represent the positive effect of the sample on the prediction, and negative values represent the negative effect. Crowded areas indicate a large number of samples gathered together. Color indicates the size of values: red indicates high feature values and blue indicates low feature values. Turbidity shows a positive correlation with SHAP value for the iron prediction. The impact on Fe prediction is greatly affected by large turbidity values with SHAP values obviously increase. In addition, the impact of hardness and TDS are similar to turbidity, but a portion of their samples are concentrated around the 0 value of SHAP,

so they have much less impact on the prediction of Fe. The variation of pH and conductivity barely affects the prediction of Fe, as such the majority of samples clustered around the 0 value of SHAP.

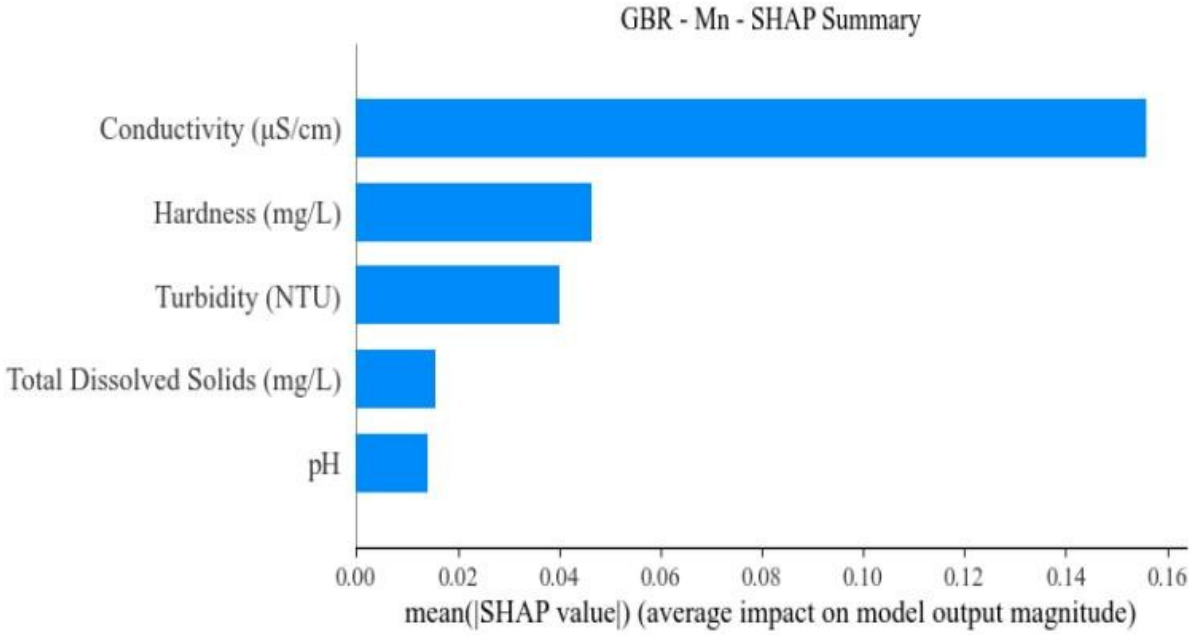


Figure 4.12 SHAP summary plot of Mn of GBR model

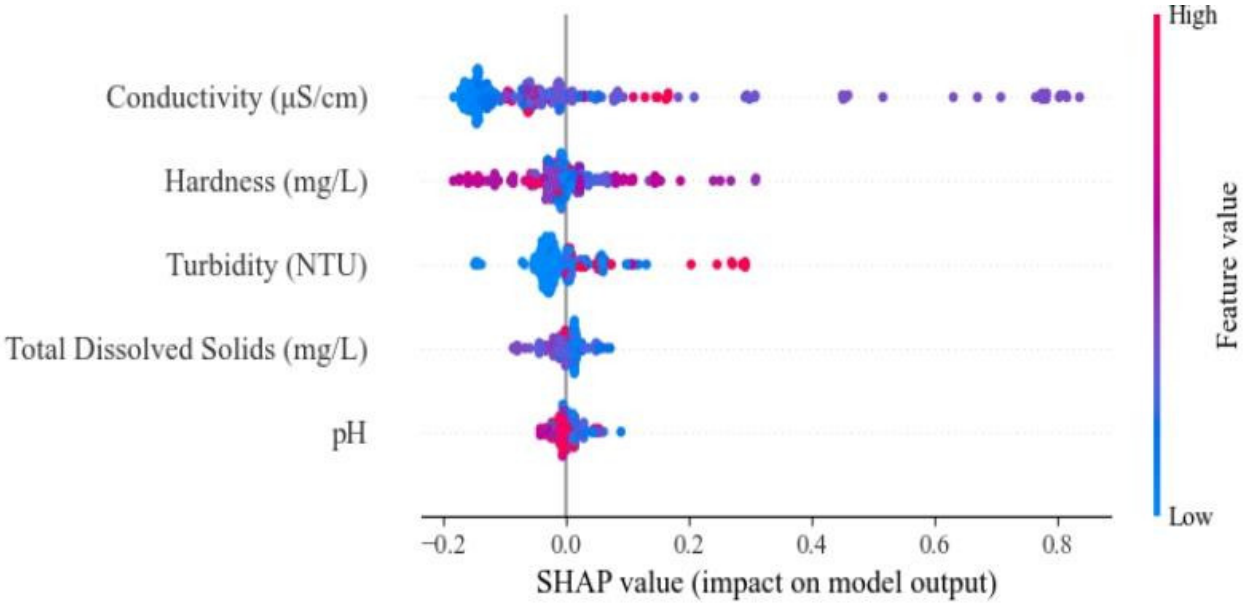


Figure 4.13 SHAP scatter diagram of Mn of GBR model

Conductivity is the least important factor in the prediction of Fe, conversely, it shows the most important effect for Mn prediction as shown in Figure 4.12. Then followed by hardness, turbidity, TDS and pH. Among these features, pH and TDS tend to have less effects to the prediction of both Fe and Mn. From Figure 4.13, we can see the values in the middle (purple color) among all the conductivity values tend to have a positive correlation regarding the impact of Mn prediction. As for the hardness, the lower the value, the less the effect on Mn prediction. Most of the negative turbidity values cluster on the negative side of the SHAP value, which means the lower the turbidity value, the more likely it is to show a negative correlation. Most of the TDS and pH values concentrate around 0 SHAP values, indicating the effect to Mn prediction is low.

4.2 GAN Based Modelling

Single-parameter data augmentation was applied to GAN data augmentation method: the process of generating synthetic data of iron and manganese were analyzed separately.

4.2.1 Statistical analysis

4.2.1.1 Correlation analysis of iron

As Figure 4.14 shows, the correlation coefficients of iron and turbidity is 0.83, showing the strongest correlation in this heatmap. pH shows some correlations with iron and turbidity, with the correlation coefficients are 0.4 and 0.47, respectively.

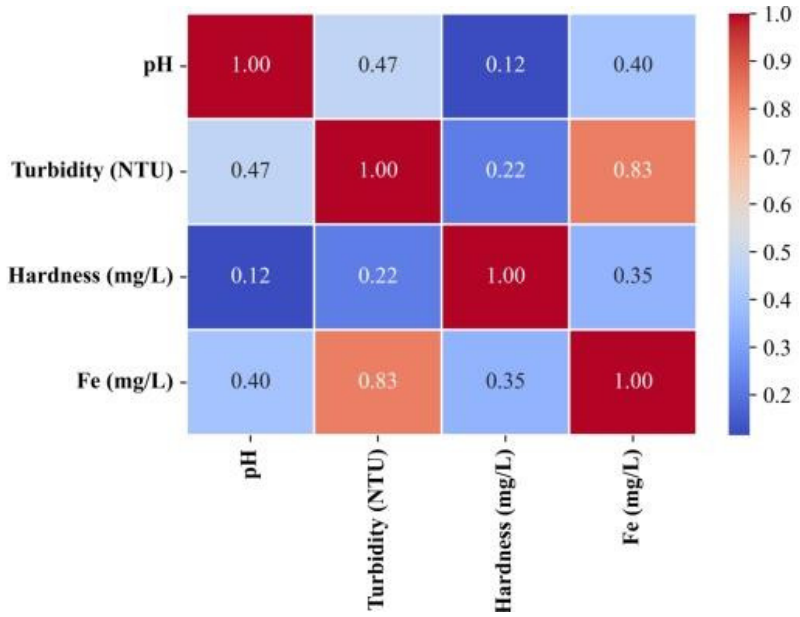


Figure 4.14 Pearson Correlation Heatmap of Fe in GAN

4.2.1.2 Correlation analysis of manganese

Compared with iron, in the Pearson Correlation Heatmap of manganese, the correlation between every parameter is not very significant. Turbidity and pH show the same correlation as in Figure 4.14, with the correlation coefficient is 0.47.

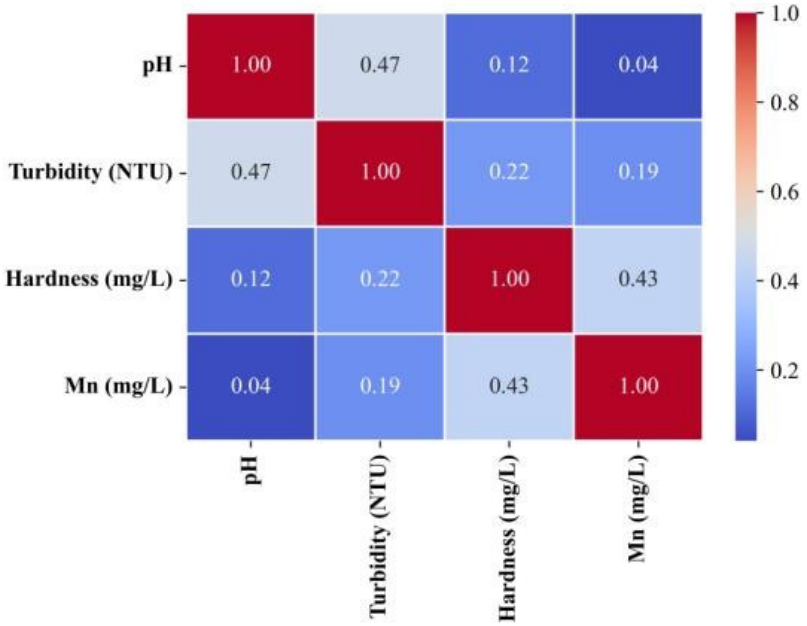


Figure 4.15 Pearson Correlation Heatmap of Mn in GAN

4.2.2 GAN sample generation

4.2.2.1 GAN sample generation for iron

After the GAN training, 1920 samples were generated initially. The maximum iteration was set as 10 to acquire effective synthetic data and less time and there were 1000 samples generated each iteration. The epochs and batch size were set as 15 and 128. There were 1740 samples of them that were identified as correct samples by the Discriminator in the process of GAN samples generation. However, 1516 generated samples were deleted because of high similarity and 1 sample with negative values also needed to be withdrawn. No out-of-range samples in the generation process were found. As a result, a total of 223 sets of iron and its three predictive parameters samples remained to be discussed in the next ML process.

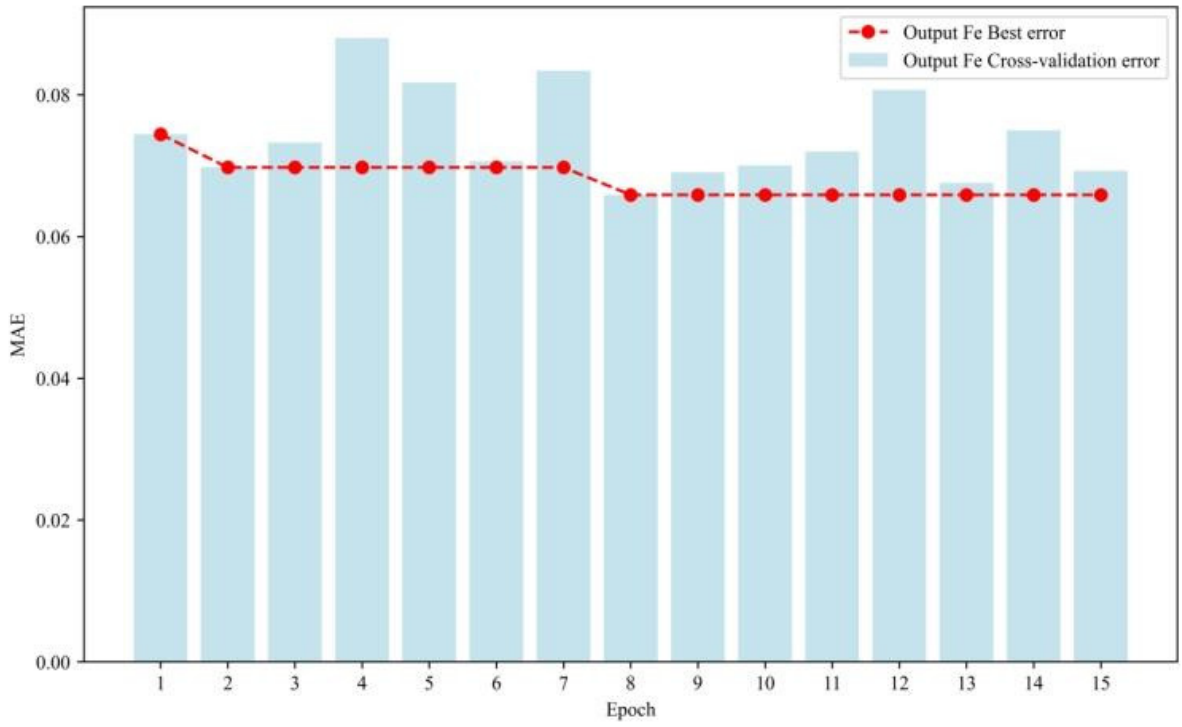


Figure 4.16 Iron-cross-validation error and best error during loop iteration

4.2.2.2 GAN sample generation for manganese

Like the process of generating iron related data, 1920 samples were generated initially. The maximum iteration was set as 10 and 1000 was set to be the generated samples each iteration. The epoch and batch size were also set as 15 and 128. There were 1079 samples of them were identified as correct samples by discriminator and 800 generated samples were deleted because of high similarity and no out-of-range samples or negative values in the generation process were found. In total, 276 sets of samples remained finally.

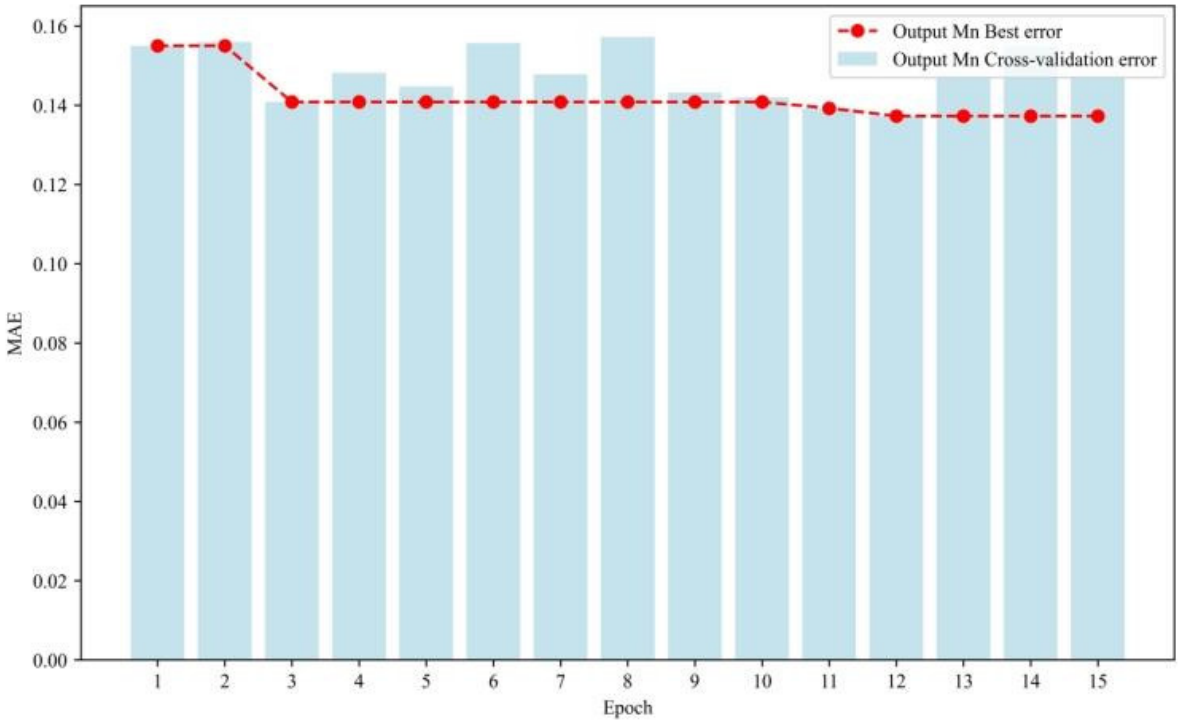


Figure 4.17 Manganese-cross-validation error and best error during loop iteration

Figures 4.16 and 4.17 are the demonstrations of iron-cross-validation error and best error during loop iteration and manganese-cross-validation error and best error during loop iteration. These two figures show similar trends. The best cross-validation error of the synthetic data for both figures slightly reduced during the process of iterations and loops,

which indicates the improvement of the augmented data quality. In addition, the cross-validation error overall remains at a similar level at the end. A possible reason could be the GAN model had leaned the characteristics of the data and reached the apex of performance.

4.2.3 Model results

4.2.3.1 Model results of iron prediction

Four ML methods were utilized to simulate iron prediction, with the GAN augmented data. Performance results and best parameters for iron prediction of each model are shown in table 4.4, and the visualized model results are shown as scatter figures in Figure 4.18.

From the results of RF, XGB, GBR, and DT models, the GBR model shows the best performance in predicting iron concentration, with the train R^2, test R^2, train RMSE, and Test RMSE score at 0.999, 0.994, 0.002, and 0.037, respectively. Besides, RF, XGB and GBR models all achieved great performance: the train and test R^2 of these 3 models are greater than 0.99, indicating a great fitting ability; the train and test RMSE are all less than 0.04, reflecting small errors between predicted samples and actual samples. From the scatter figures 4.18, most of the training and testing data fell very close to the regression line.

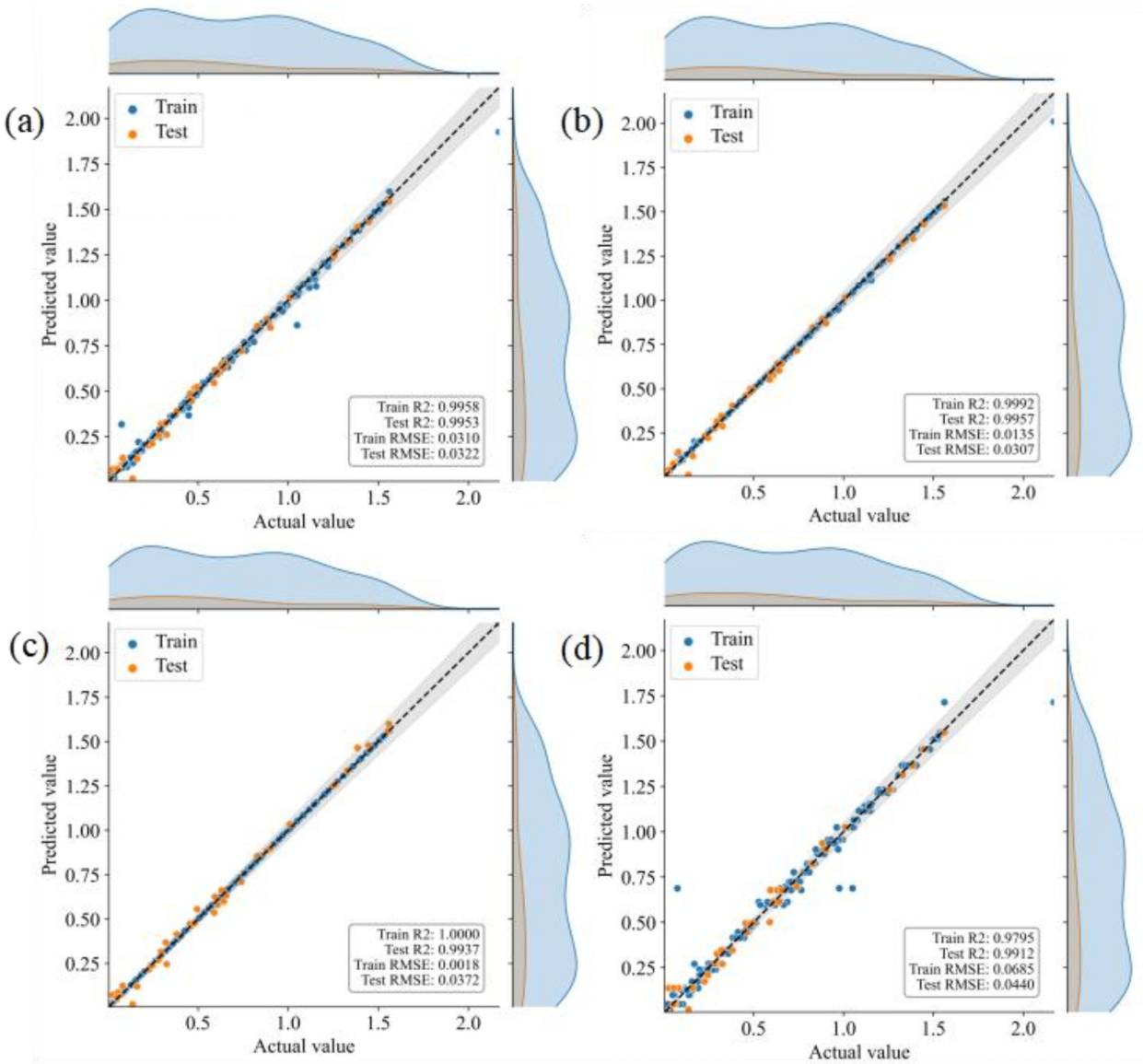


Figure 4. 18 Scatter regression plots of (a) RF; (b) XGB; (c) GBR; (d) DT for Fe prediction

Table 4.4 Performance of GAN based models and best hyperparameters for Fe prediction

Model	Best Parameters	Train R^2	Test R^2	Train RMSE	Test RMSE
RF	{'max_depth': 40, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}	0.996	0.996	0.029	0.030
XGB	{'learning_rate': 0.05, 'max_depth': 10, 'n_estimators': 100, 'subsample': 0.9}	0.999	0.996	0.014	0.031
GBR	{'learning_rate': 0.1, 'max_depth': 40, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}	1	0.994	0.002	0.037
DT	{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 5}	0.980	0.991	0.068	0.043

4.2.3.2 Model results of manganese prediction

For manganese prediction using GAN augmented data, the GBR model has the best performance, same as the results for iron prediction. The train R^2, test R^2, train RMSE, and test RMSE all acquired the best scores among these 4 models, which are 0.9995, 0.988, 0.005 and 0.028, reflecting excellent simulation process and accurate prediction for manganese. However, RF didn't show great simulation this time compared to its other performances in this study, with test R^2 is only 0.819 and test RMSE is greater than 0.1.

Table 4.5 Performance of GAN based models and best hyperparameters for Mn prediction

Model	Best Parameters	Train R^2	Test R^2	Train RMSE	Test RMSE
RF	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 10}	0.936	0.819	0.055	0.108
XGB	{'learning_rate': 0.1, 'max_depth': 20, 'n_estimators': 20, 'subsample': 0.8}	0.941	0.929	0.053	0.068
GBR	{'learning_rate': 0.1, 'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100}	0.9995	0.988	0.005	0.028
DT	{'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5}	0.992	0.975	0.020	0.040

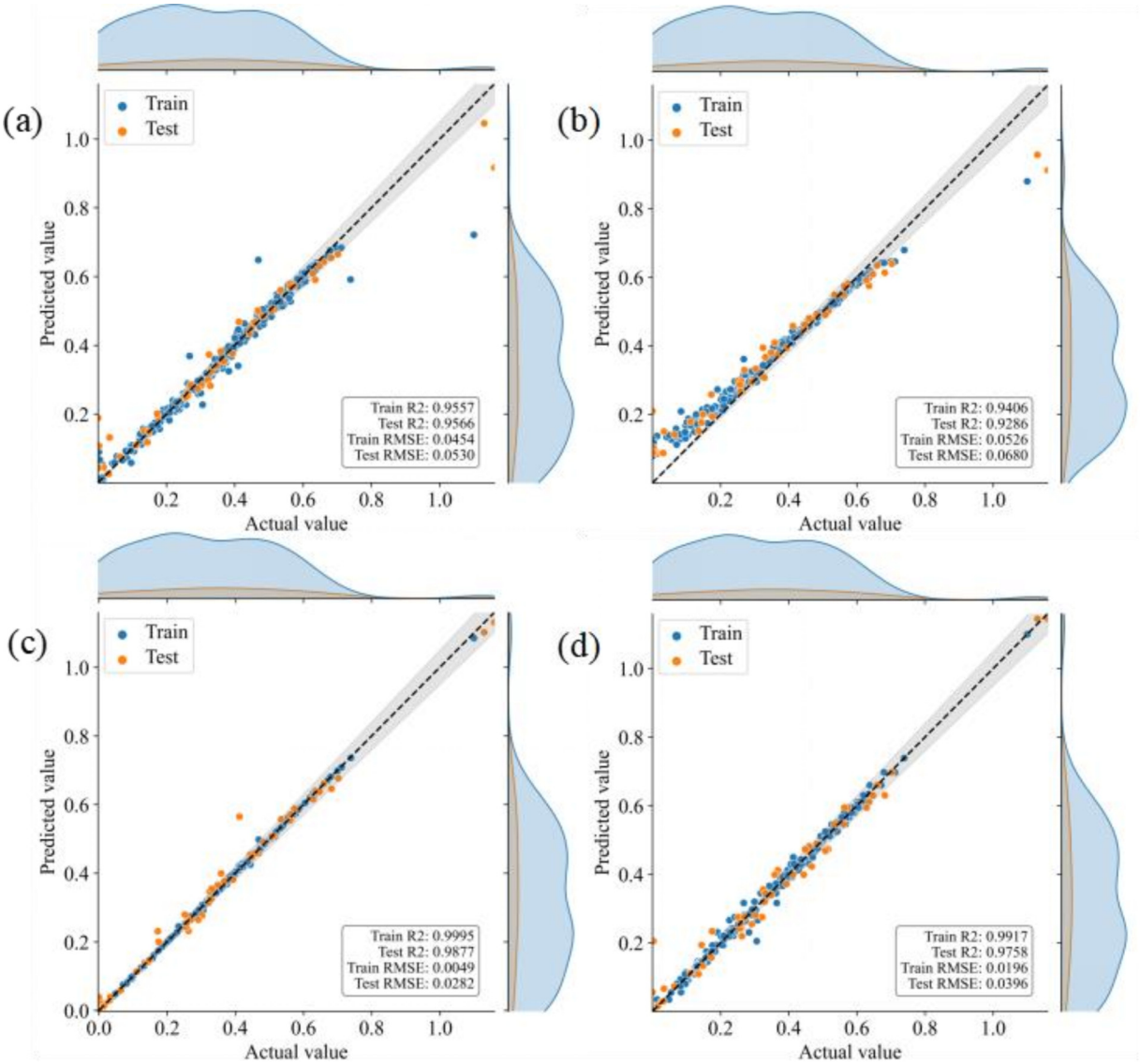


Figure 4.18 Scatter regression plots of (a)RF; (b)XGB; (c) GBR; (d) DT for Mn prediction

4.2.4 Interpretable analysis

4.2.4.1 Feature importance

Feature importance ranking will be presented by the best model GBR in predicting both iron and manganese after GAN augmentation. From Figure 4.20, turbidity is the only significant impact on the iron prediction, with the importance scores close to 1. The effect

of pH and hardness is negligible. As for Mn prediction, hardness is the most important feature, followed by turbidity and pH, showing some impact to Mn prediction, but scored below 0.2 in both.

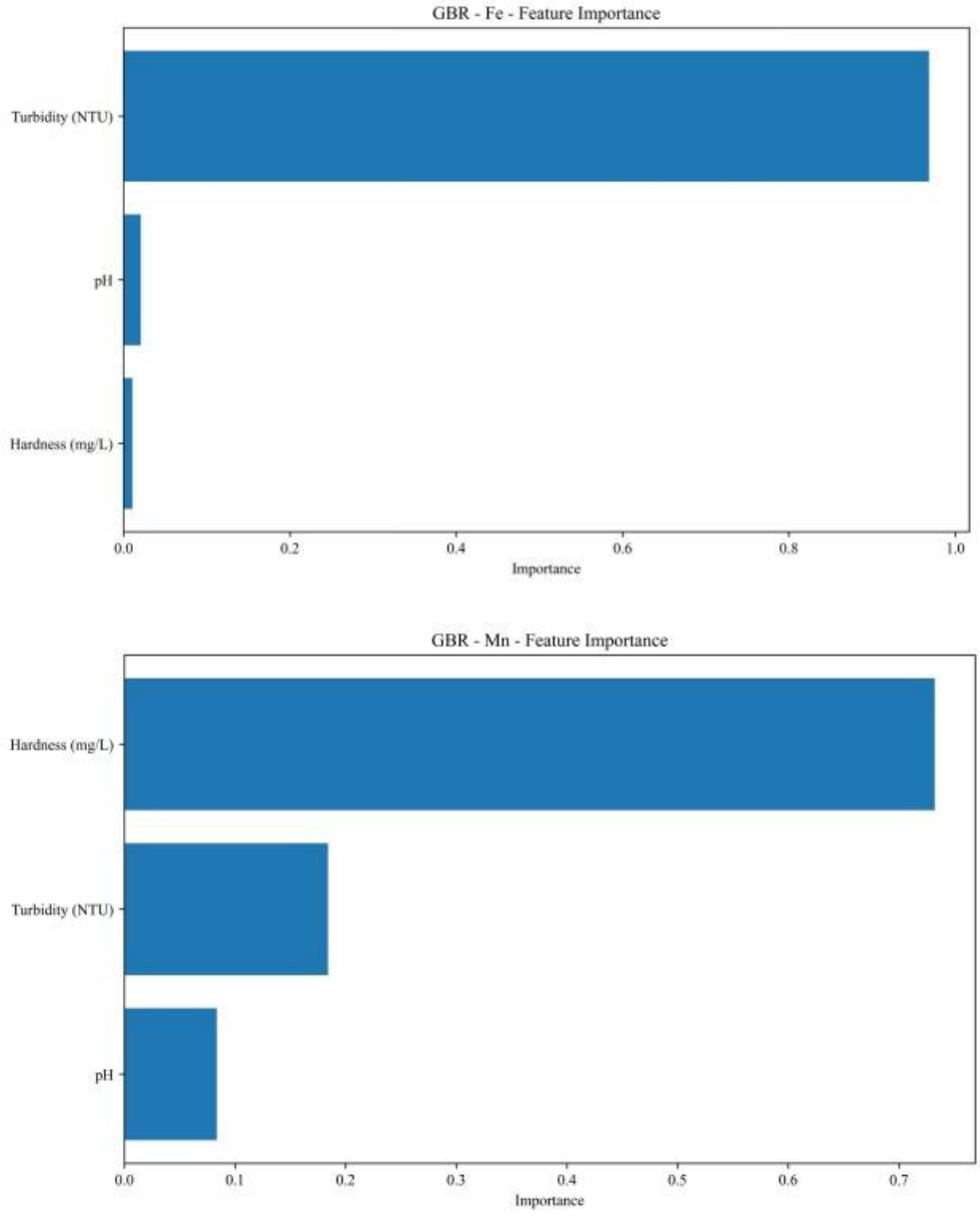


Figure 4.19 Feature importance of GBR model after GAN augmentation

4.2.4.2 Partial importance

The dependence of pH, turbidity and hardness to iron and manganese prediction will be discussed below. Take the best model, GBR as an example. As shown in Figure 4.21, the predicted iron concentration increases significantly with the increase of turbidity, correspondent to the feature importance. However, pH and hardness didn't show obvious partial dependence to iron prediction. In the manganese partial dependence plots, there is a small negative correlation between pH and manganese; with the increase of turbidity, the manganese concentration fluctuates suddenly only to drop down when turbidity is around 2 NTU and then increases gradually. Moreover, the predicted manganese content has a dramatic positive correlation with hardness ranging from 0 to 400 mg/L.

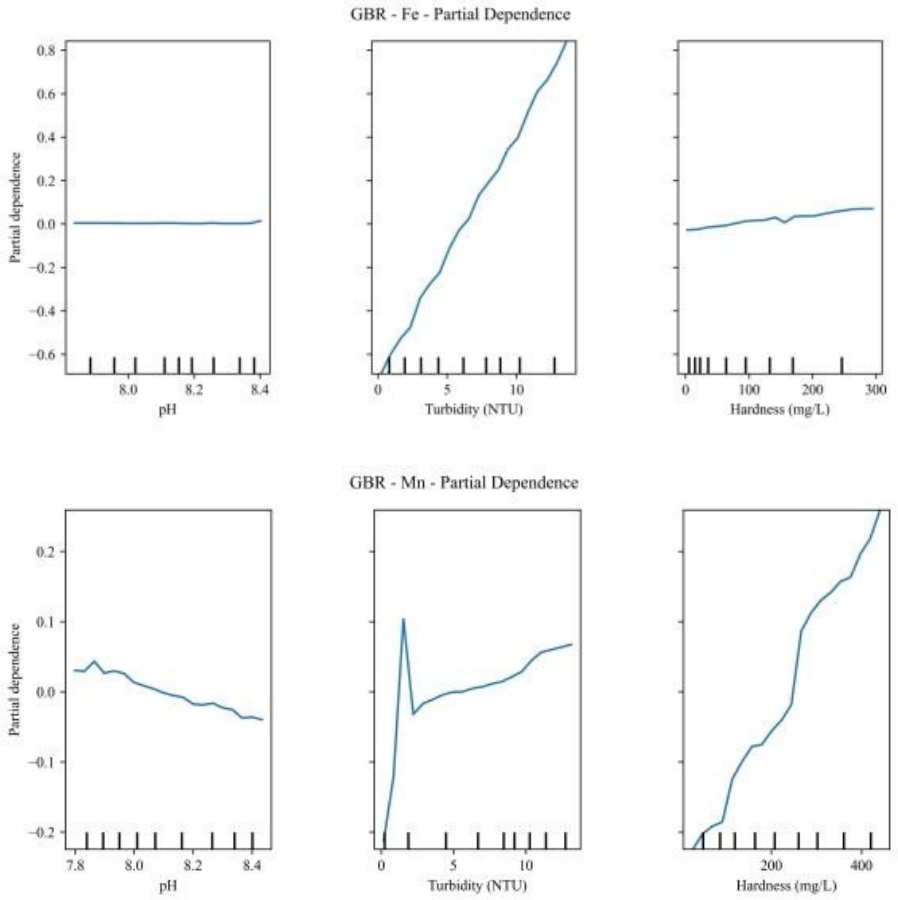


Figure 4.20 Partial Dependence Plots of GBR model

Figure 4.22 indicates the two-way partial dependence correlations based on GBR model. The common influence through pH and turbidity to the iron concentration is regular, this is shown by the increase of iron concentration of around 0.2mg/L for every 2 units of turbidity. The combined effect of hardness and turbidity also shows a similar pattern: with the change of hardness or pH, the concentration of iron has very limited reflections. As for manganese, the maximum concentration happens when turbidity is about 2.4 NTU and hardness is about 260-300 mg/L. The combined effect of pH and turbidity or pH and hardness doesn't show significant effect to manganese prediction.

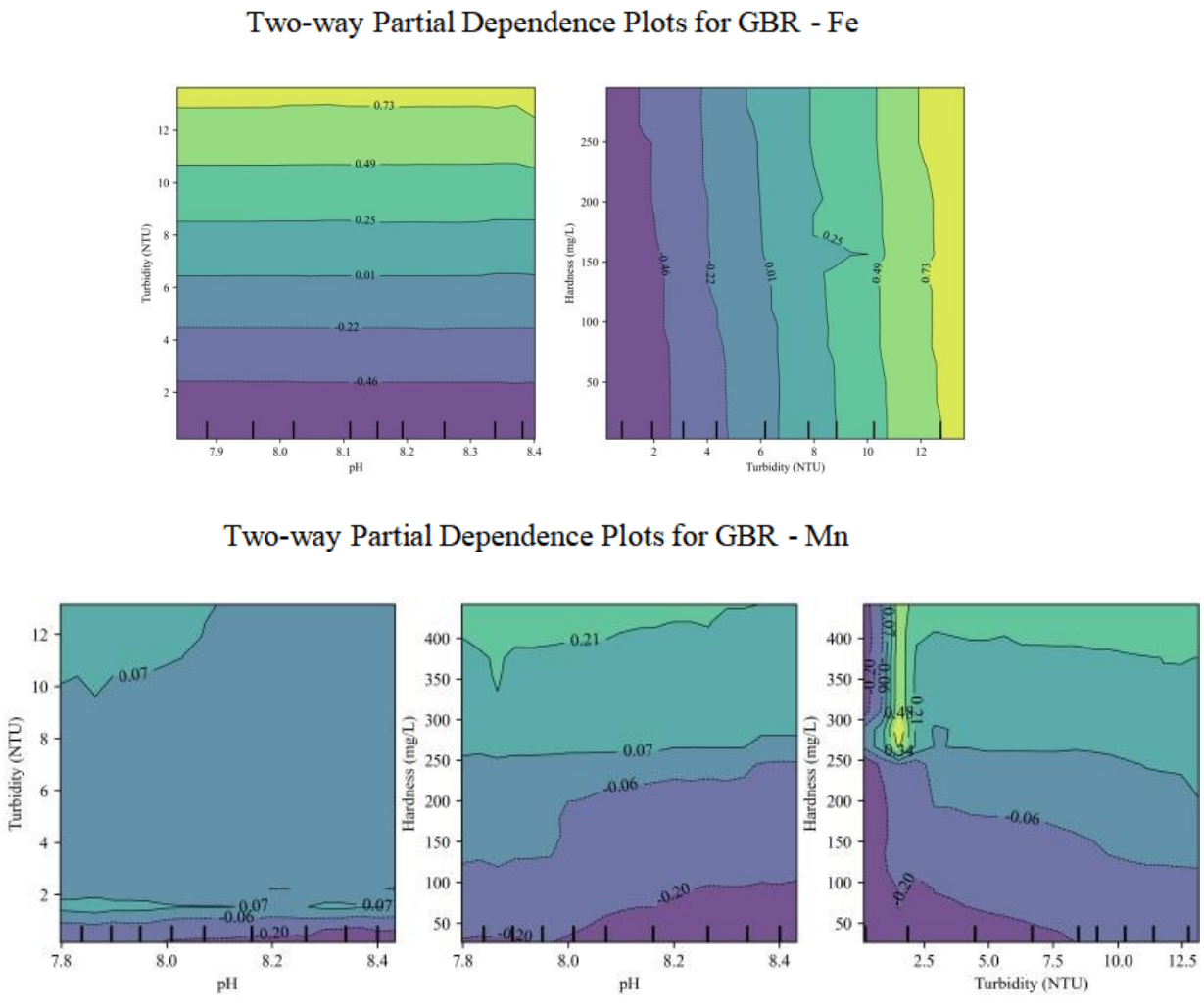


Figure 4.21 Two-way Partial Dependence Plots of GBR model

4.2.4.3 SHAP analysis

Figure 4.23 represents the average impact size of pH, turbidity and hardness on iron and manganese prediction ranked by SHAP values. For manganese prediction, hardness is the most significant effecting factor, followed by turbidity and pH. Turbidity is the most important factor affecting the iron amount and the predicted effect of pH and hardness on iron shows limited impact represented by SHAP values less than 0.04.

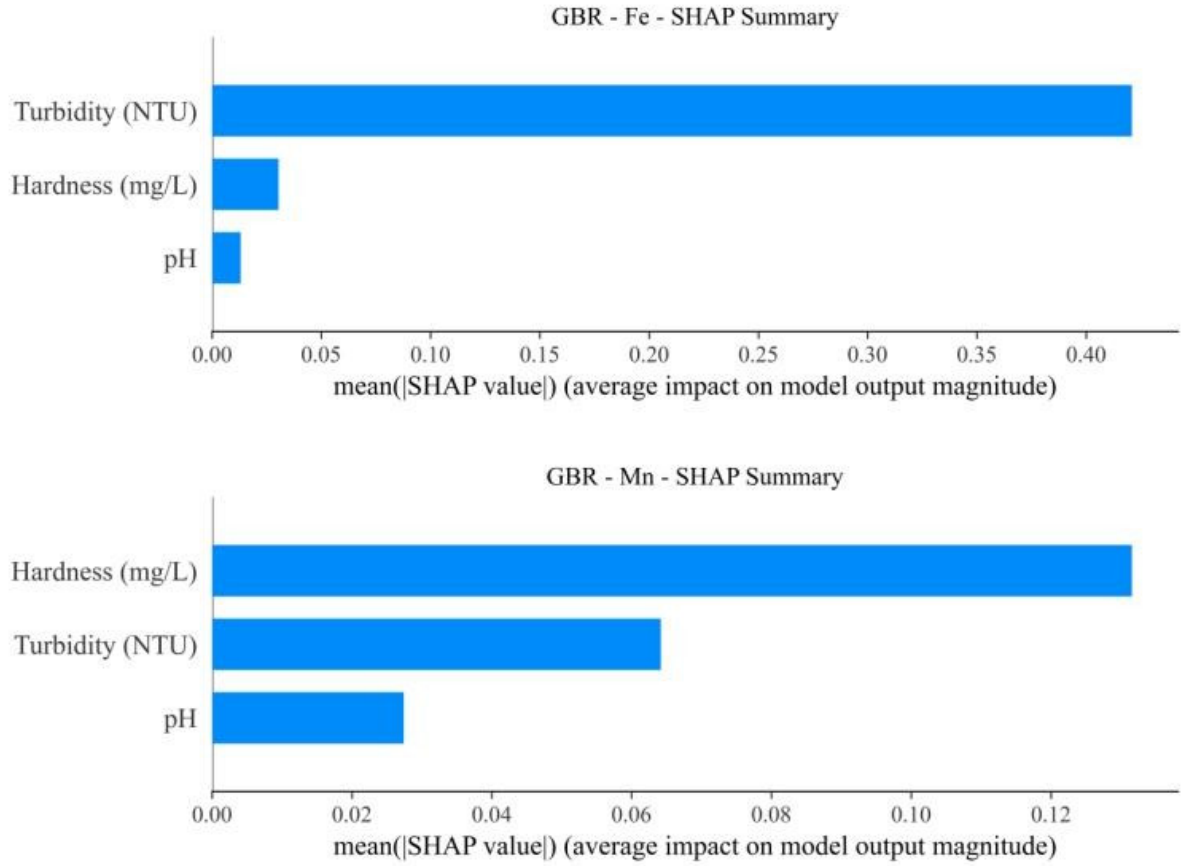


Figure 4.22 SHAP summary plot of GBR model after GAN

Figure 4.24 represents the SHAP scatter diagrams of iron and manganese SHAP analysis. There is absolute evidence of a positive correlation between turbidity and SHAP

values on iron prediction, represented with the increase of the amounts of turbidity, the impact of turbidity on iron output shifts from negative to positive gradually. Hardness also shows a similar positive correlation, however, compared with turbidity, the influence level is much smaller, and a large amount of hardness samples gathered around 0, which means those samples have little impact on iron prediction. Same as pH, most of the samples have negligible effect but only some large readings of pH show some impact on iron output.

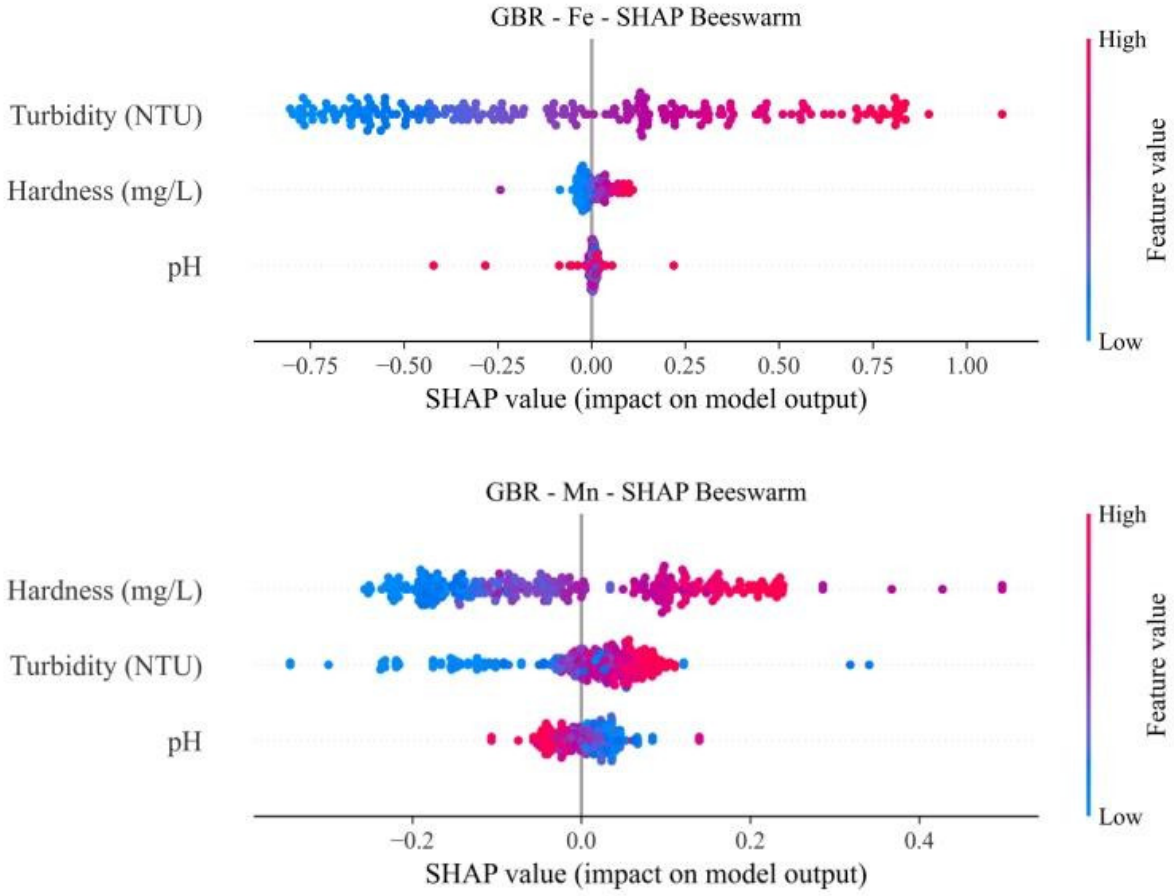


Figure 4.23 SHAP scatter diagram of GBR model after GAN

As shown in manganese SHAP scatter diagram, hardness and turbidity indicate positive correlations with SHAP values while pH shows a very limited negative correlation. Additionally, most of the turbidity samples clustered around 0-0.1 on the x-axis and some

low values (blue points) fell on the left side of the horizontal axis, indicating most of them have a minor impact on model output; however, some small pH values have a negative impact in predicting manganese.

4.3 Graphical user interface

Prediction application was developed to be utilized in reality for iron and manganese prediction. Figures 4.25 and 4.26 represent two graphical user interfaces of applications utilized two different technologies explored in this study. This process was completed on MATLAB software.

The graphical user interface based on GAN technology shows more application prospects. There are two main reasons: (1) Only 3 input parameters need to be measured ahead to predict iron and manganese concentrations; (2) Compared to traditional data augmentation methods, GAN shows it powerful diverse synthetic samples generation ability, given its AI algorithm.

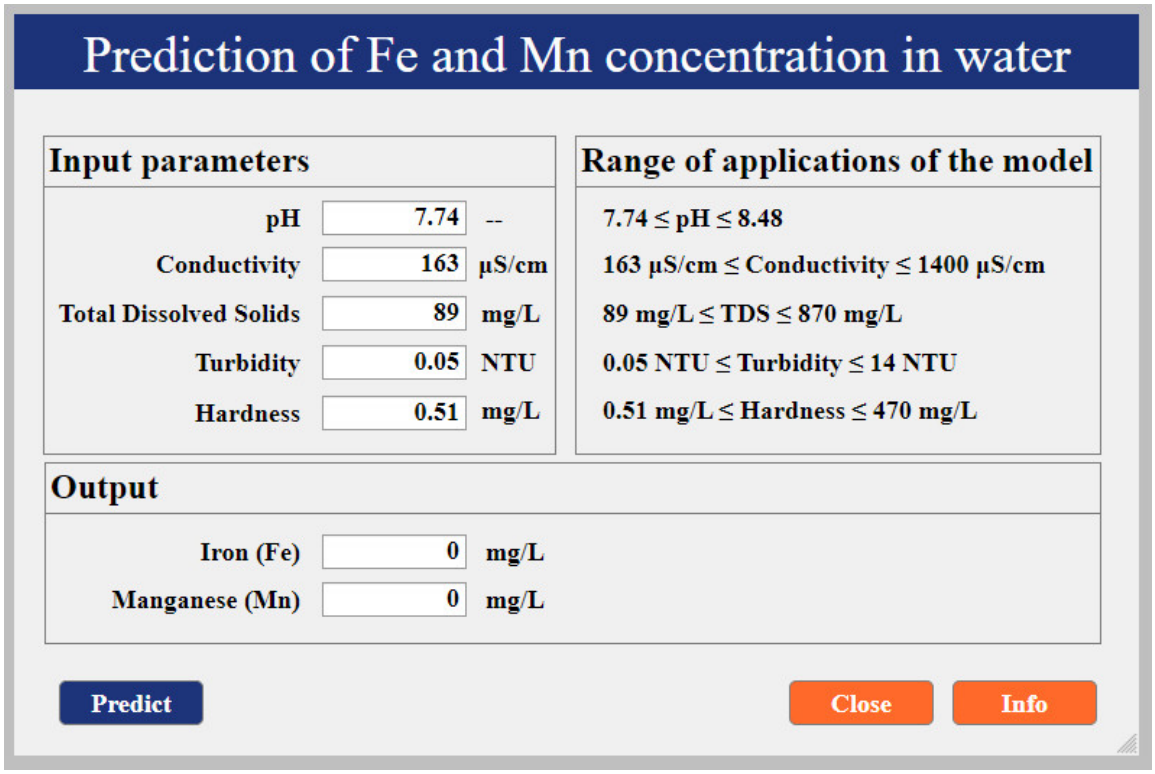


Figure 4.24 Graphical user interface with 5 parameters

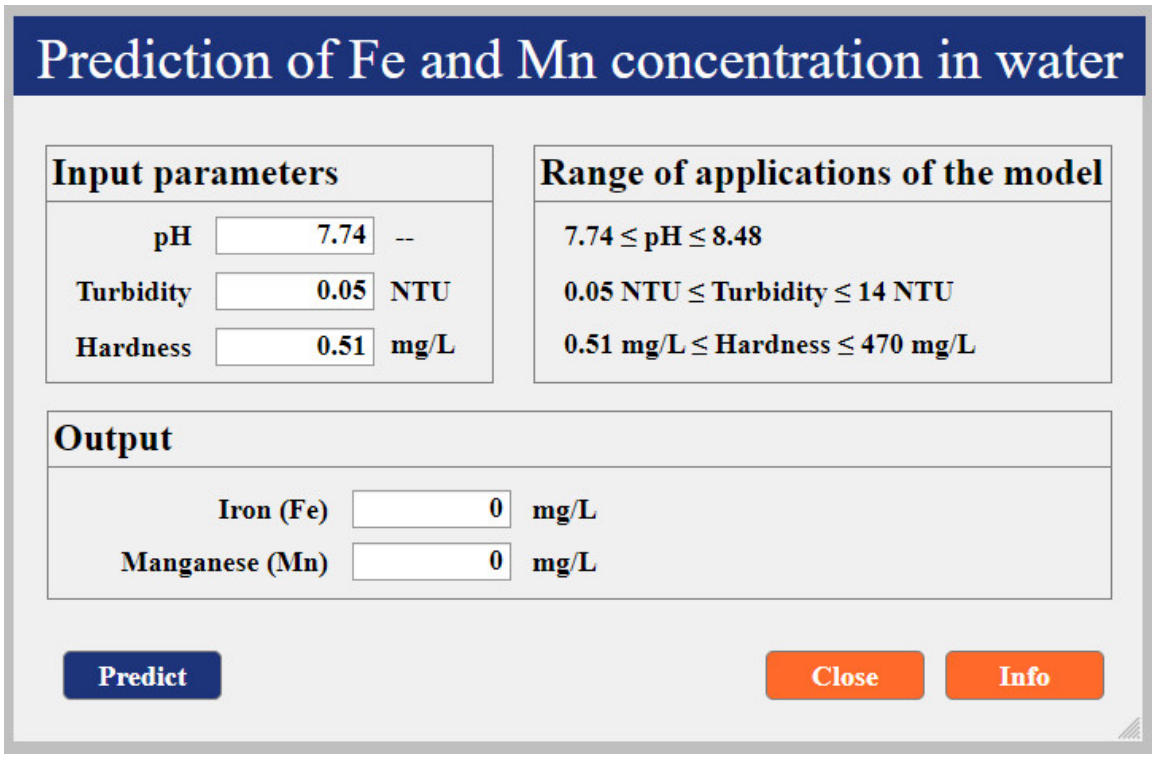


Figure 4.25 Graphical user interface with 3 parameters

Chapter 5 Conclusions

5.1 Research summary

In this study, the drinking water prediction models of Iron and Manganese concentration in BC's small, rural and remote First Nations communities were investigated using ML. Multiple data augmentation methods were developed to acquire effective synthetic data. Five physical indexes which are TDS, conductivity, pH, turbidity, and hardness were selected to serve as predictive values, namely input parameters. The model results of 4 ML algorithms were compared to selecting the best optimum model to develop the prediction interface for the visualization. Statistical analysis and interpretable analysis were also conducted to collect information on the correlation and importance of the predicted values. The main findings of this study are as follows:

(1) According to the trend statistics, the measures of dispersion, the distribution statistics analysis, and cross-validation, Numerical Interpolation-Bootstrapping-Noise Injection (NI-BS-NI) based data augmentation and GAN based data augmentation both synthesized reliable data for modelling.

(2) Two models based on different augmented data and predictors were developed in this study. Considering GAN based model the better one to be applied in the prediction tool.

(3) The GBR model performed the best for both NI-BS-NI based modelling and GAN based modelling in predicting iron and manganese. The Train R^2 of two models are just neath of 1, and Test R^2 achieved very ideal results. All the RMSE scores are below 0.06. The ideal evaluation results prove the excellence and effectiveness of the GBR model.

(4) Iron and turbidity show great correlation and have little correlation with other features. In addition, manganese concentration shows a significant positive correlation with conductivity.

(5) For the iron prediction, turbidity has an outstanding impact, with the importance scores over 0.8. Followed by hardness, TDS, conductivity, and pH. As for manganese prediction, conductivity is the most important feature, followed by hardness and turbidity.

5.2 Limitations and future research

In this study, the prediction of Fe and Mn concentration in drinking water of First Nations SRR areas in BC using ML was investigated to access Fe and Mn concentration on site rapidly, saving the time and costs needed to send them to laboratory tests. Although this attempt demonstrated ideal models during experiments, its feasibility of predicting Fe and Mn in practical field applications has not been verified. There are still some steps we can take to make this method really applied for the actual water quality detection. Recommendations for the follow-up optimization and possible future studies are listed as follows:

(1) Due to the very limited water quality data, only 2 heavy metal pollutants, iron and manganese were investigated in this study. In the future, more water quality data should be collected, combined with data augmentation methods, more kinds of pollutants should be involved to achieve fast and accurate detection. The prediction is not limited to only heavy metal pollutants, other hard-to-get pollutants, such as E. coli, organic matters, can also be predicted.

(2) The prediction method developed in this study needs to be verified in actual presence. The predicted values acquired by using the decision tool should be compared with true values measured in lab to verify its effectiveness and reliability.

(3) Turbidity, which is the one predictive parameter in this study, is dependent on the form of iron. The water samples were collected from the groundwater, containing more ferrous iron (Fe^{2+}) due to limited oxygen. While travelling to the lab, ferrous iron was oxidized to ferric iron (Fe^{3+}) due to exposure to air. Then the solubility in the water samples decreased, affecting turbidity values. In this study, the models were established based on the laboratory water samples data. However, in actual use, all input index data will be obtained immediately by field meters, so it needs to be explored whether there is an impact on the accuracy of the prediction results in practical applications.

(4) In GAN data augmentation process, the cross-validation error didn't show an obvious reduction in the trend could also be caused by insufficient data or the inherent complexity of the data. This will be clarified if more raw sampling data can be acquired.

(5) In future study, adding and adjusting different input parameters for modelling is worthy of consideration. The reason is unknown correlations between different physical or chemical parameters might help build up better models due to the interactive correlations between input values and predicted values.

References

Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermalenergy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production, 203*, 810-821.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water, 11*(11), 2210.

Al-Razee, A., Abser, M. N., Mottalib, M. A., Rahman, M. S., & Cho, N. (2019). Assessment of heavy metals in sediments of Shitalakhya River, Bangladesh. *분석과학, 32*(5), 210-216.

Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics and Biomechanics, 2020*, 6659314.

Alias, N., Rosli, S. A., Sazalli, N. A. H., Hamid, H. A., Arivalakan, S., Umar, S. N. H., . . . Lockman, Z. (2020). 15 - Metal oxide for heavy metal detection and removal. In Y. Al-Douri (Ed.), *Metal Oxide Powder Technologies* (pp. 299-332): Elsevier.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 82*(4), 1059-1086.

Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of environmental chemical engineering, 9*(1), 104599.

Ayangbenro, A. S., & Babalola, O. O. (2017). A New Strategy for Heavy Metal Polluted Environments: A Review of Microbial Biosorbents. *International Journal of Environmental Research and Public Health, 14*(1), 94.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment, 8*(2), 2793-2801.

Babbar, R., & Babbar, S. (2017). Predicting river water quality index using data mining techniques. *Environmental Earth Sciences, 76*, 1-15.

Baijius, W., & Patrick, R. J. (2019). “We don’t drink the water here”: the reproduction of undrinkable water for First Nations in Canada. *Water, 11*(5), 1079.

Balasooriya, B. K., Rajapakse, J., & Gallage, C. (2023). A review of drinking water quality issues in remote and Indigenous communities in rich nations with special emphasis on Australia. *Science of the Total Environment*, 166559.

Baptista, M. L., Goebel, K., & Henriques, E. M. (2022). Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artificial Intelligence, 306*, 103667.

Briffa, J., Sinagra, E., & Blundell, R. (2020). Heavy metal pollution in the environment and their toxicological effects on humans. *Heliyon, 6*(9), e04691.

British Columbia Assembly of First Nations. (2023). Lheidli T'enneh First Nation. https://www.bcafn.ca/first-nations-bc/cariboo/lheidli-tenneh-first-nation

Brown, J., Acey, C. S., Anthonj, C., Barrington, D. J., Beal, C. D., Capone, D., . . . Hicks, B. (2023). The effects of racism, social exclusion, and discrimination on achieving universal safe water and sanitation in high-income countries. *The Lancet Global Health, 11*(4), e606-e614.

Bua, D. G., Annuario, G., Albergamo, A., Cicero, N., & Dugo, G. (2016). Heavy metals in aromatic spices by inductively coupled plasma-mass spectrometry. *Food Additives & Contaminants: Part B, 9*(3), 210-216.

Chen, Y., Zhu, Y., & Chang, Y. (2020). *CycleGAN Based Data Augmentation For Melanoma images Classification*. Paper presented at the Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition, Xiamen, China.

Chowdhury, S., Champagne, P., & McLellan, P. J. (2009). Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. *Science of the Total Environment, 407*(14), 4189-4206.

Connor, S., Khoshgoftaar, T. M., & Borko, F. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*(1).

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Daley, K., Jamieson, R., Rainham, D., & Truelstrup Hansen, L. (2018). Wastewater treatment and public health in Nunavut: a microbial risk assessment framework for the Canadian Arctic. *Environmental Science and Pollution Research, 25*(33), 32860-32872.

Deng, Z., Yang, Z., Ma, X., Tian, X., Bi, L., Guo, B., . . . Zhang, S. (2018). Urinary metal and metalloid biomarker study of Henoch-Schonlein purpura nephritis using inductively coupled plasma orthogonal acceleration time-of-flight mass spectrometry. *Talanta, 178*, 728-735.

Dezfooli, D., Hosseini-Moghari, S.-M., Ebrahimi, K., & Araghinejad, S. (2018). Classification of water quality status based on minimum quality parameters: application of machine learning techniques. *Modeling Earth Systems and Environment, 4*, 311-324.

Duignan, S., Moffat, T., & Martin-Hill, D. (2022). Be like the running water: Assessing gendered and age-based water insecurity experiences with Six Nations First Nation. *Social Science & Medicine, 298*, 114864.

El Bilali, A., & Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. *Journal of the Saudi Society of Agricultural Sciences, 19*(7), 439-451.

Fekri, M. N., Ghosh, A. M., & Grolinger, K. (2019). Generating energy data for machine learning with recurrent generative adversarial networks. *Energies, 13*(1), 130.

Fernández-Martínez, R., Rucandio, I., Gómez-Pinilla, I., Borlaf, F., García, F., & Larrea, M. T. (2015). Evaluation of different digestion systems for determination of trace mercury in seaweeds by cold vapour atomic fluorescence spectrometry. *Journal of Food Composition and Analysis, 38*, 7-12.

First Nations Health Authority. (2023). Monthly Drinking Water Advisories in First Nations Communities in BC - November 2023. https://www.fnha.ca/Documents/Drinking-Water-Advisory-Monthly-Summary.pdf

Government of Canada. (2017). First Nations People in Canada. https://www.rcaanc-cirnac.gc.ca/eng/1307460755710/1536862806124

Government of Canada. (2021). Indigenous peoples and communities. https://www.rcaanc-cirnac.gc.ca/eng/1100100013785/1529102490303

Government of Canada. (2023). Ending long-term drinking water advisories. https://www.sac-isc.gc.ca/eng/1506514143353/1533317130660

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics, 24*(1), 44-65.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144.

Gumpu, M. B., Sethuraman, S., Krishnan, U. M., & Rayappan, J. B. B. (2015a). A review on detection of heavy metal ions in water–an electrochemical approach. *Sensors and Actuators B: Chemical, 213*, 515-533.

Gumpu, M. B., Sethuraman, S., Krishnan, U. M., & Rayappan, J. B. B. (2015b). A review on detection of heavy metal ions in water – An electrochemical approach. *Sensors and Actuators B: Chemical, 213*, 515-533.

Health Canada. 2019. Guidelines for Canadian Drinking Water Quality.

https://publications.gc.ca/collections/collection_2019/sc-hc/H144-13-14-2019-eng.pdf

Health Canada. 2021. Guidelines for Canadian Drinking Water Quality - Summary Tables.

https://www.canada.ca/en/health-canada/services/environmental-workplace-

health/reports-publications/water-quality/guidelines-canadian-drinking-water-quality-

summary-table.html

Hu, G., Mian, H. R., Abedin, Z., Li, J., Hewage, K., & Sadiq, R. (2022). Integrated probabilistic-fuzzy synthetic evaluation of drinking water quality in rural and remote communities. *Journal of Environmental Management, 301*, 113937.

Hu, G., Mian, H. R., Mohammadiun, S., Rodriguez, M. J., Hewage, K., & Sadiq, R. (2023). Appraisal of machine learning techniques for predicting emerging disinfection byproducts in small water distribution networks. *Journal of hazardous materials, 446*, 130633.

Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). *Global explanations of neural networks: Mapping the landscape of predictions.* Paper presented at the Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.

Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., & Gómez-Canaval, S. (2022). Data augmentation techniques in time series domain: A survey and taxonomy. *arXiv preprint arXiv:2206.13508*.

Imani, M., Hasan, M. M., Bittencourt, L. F., McClymont, K., & Kapelan, Z. (2021). A novel machine learning application: Water quality resilience prediction Model. *Science of the Total Environment, 768*, 144459.

Indigenous Foundations. (2009). Aboriginal Identity & Terminology. https://Indigenousfoundations.arts.ubc.ca/aboriginal_identity__terminology/

Inoue, H. (2018). Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*.

Islam, M., & Yuan, Q. (2018). First Nations wastewater treatment systems in Canada: Challenges and opportunities. *Cogent Environmental Science, 4*(1), 1458526.

Jaishankar, M., Tseten, T., Anbalagan, N., Mathew, B. B., & Beeregowda, K. N. (2014). Toxicity, mechanism and health effects of some heavy metals. *Interdiscip Toxicol, 7*(2), 60-72.

Khan, Y., & See, C. S. (2016). *Predicting and analyzing water quality using machine learning: a comprehensive model.* Paper presented at the 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT).

Kim, H., Harrison, F. E., Aschner, M., & Bowman, A. B. (2022). Exposing the role of metals in neurological disorders: a focus on manganese. *Trends in Molecular Medicine, 28*(7), 555-568.

Kingsbury, B. (1998). "Indigenous peoples" in international law: a constructivist approach to the Asian controversy. *American Journal of International Law, 92*(3), 414-457.

Le Bot, B., Lucas, J.-P., Lacroix, F., & Glorennec, P. (2016). Exposure of children to metals via tap water ingestion at home: Contamination and exposure data from a nationwide survey in France. *Environment International, 94*, 500-507.

Li, P., Karunanidhi, D., Subramani, T., & Srinivasamoorthy, K. (2021). Sources and consequences of groundwater contamination. *Archives of environmental contamination and toxicology, 80*, 1-10.

Li, X., Yang, Y., Yang, J., Fan, Y., Qian, X., & Li, H. (2021). Rapid diagnosis of heavy metal pollution in lake sediments based on environmental magnetism and machine learning. *Journal of hazardous materials, 416*, 126163.

Lim, A. P., & Aris, A. Z. (2014). A review on economically adsorbents on heavy metals removal in water and wastewater. *Reviews in Environmental Science and Bio/Technology, 13*, 163-181.

Lin, L., Wang, F., Xie, X., & Zhong, S. (2017). Random forests-based extreme learning machine ensemble for multi-regime time series prediction. *Expert Systems with Applications, 83*, 164-176.

Liu, Y., Wang, T., & Chu, F. (2024). Hybrid machine condition monitoring based on interpretable dual tree methods using Wasserstein metrics. *Expert Systems with Applications, 235*, 121104.

Lu, H., Li, H., Liu, T., Fan, Y., Yuan, Y., Xie, M., & Qian, X. (2019). Simulating heavy metal concentrations in an aquatic environment using artificial intelligence models and physicochemical indexes. *Science of the Total Environment, 694*, 133591.

Lu, S.-Y., Zhang, H.-M., Sojinu, S. O., Liu, G.-H., Zhang, J.-Q., & Ni, H.-G. (2015). Trace elements contamination and human health risk assessment in drinking water from Shenzhen, China. *Environmental Monitoring and Assessment, 187*, 1-8.

Lu, S., Zhou, Q., Ouyang, Y., Guo, Y., Li, Q., & Wang ,J. (2018). Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature Communications, 9,* 3405

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

Lyu, T., Tang, Y., Cao, H., Gao, Y., Zhou, X., Zhang, W., . . . Jiang, Y. (2023). Estimating the geographical patterns and health risks associated with PM2. 5-bound heavy metals to guide PM2. 5 control targets in China based on machine-learning algorithms. *Environmental Pollution, 337*, 122558.

Ma, Z., Wang, J., Feng, Y., Wang, R., Zhao, Z., & Chen, H. (2023). Hydrogen yield prediction for supercritical water gasification based on generative adversarial network data augmentation. *Applied Energy, 336*, 120814.

Markus, M., Tsai, C. W.-S., & Demissie, M. (2003). Uncertainty of weekly nitrate-nitrogen forecasts using artificial neural networks. *Journal of environmental engineering, 129*(3), 267-274.

Martin, Y. E., & Johnson, E. A. (2012). Biogeosciences survey:Studying interactions of the biosphere with the lithosphere, hydrosphere and atmosphere. *Progress in Physical Geography: Earth and Environment, 36*(6), 833-852.

McLeod, L., Bharadwaj, L., & Waldner, C. (2014). Risk factors associated with the choice to drink bottled water and tap water in rural Saskatchewan. *International Journal of Environmental Research and Public Health, 11*(2), 1626-1646.

McLeod, L., Bharadwaj, L. A., Daigle, J., Waldner, C., & Bradford, L. E. A. (2020). A quantitative analysis of drinking water advisories in Saskatchewan Indigenous and rural communities 2012–2016. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques, 45*(4), 345-357.

Meehan, K., Jepson, W., Harris, L. M., Wutich, A., Beresford, M., Fencl, A., . . . Wells, C. (2020). Exposing the myths of household water insecurity in the global north: A critical review. *Wiley Interdisciplinary Reviews: Water, 7*(6), e1486.

Meena, A. K., Mishra, G., Rai, P., Rajagopal, C., & Nagar, P. (2005). Removal of heavy metal ions from aqueous solutions using carbon aerogel as an adsorbent. *Journal of hazardous materials, 122*(1-2), 161-170.

Mian, H. R., Chhipi-Shrestha, G., Hewage, K., Rodriguez, M. J., & Sadiq, R. (2020). Predicting unregulated disinfection by-products in small water distribution networks: an empirical modelling framework. *Environmental Monitoring and Assessment, 192*, 1-20.

Najafzadeh, M., & Ghaemi, A. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental Monitoring and Assessment, 191*, 1-21.

Navarro-Espinoza, S., Angulo-Molina, A., Meza-Figueroa, D., López-Cervantes, G., Meza-Montenegro, M., Armienta, A., . . . Álvarez-Bajo, O. (2021). Effects of untreated drinking water at three Indigenous Yaqui towns in Mexico: insights from a murine model. *International Journal of Environmental Research and Public Health, 18*(2), 805.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika, 108*(2), 299-319.

Pang, X., Gao, T., Qiu, Y., Caffrey, N., Popadynetz, J., Younger, J., . . . Checkley, S. (2021). The prevalence and levels of enteric viruses in groundwater of private wells in rural Alberta, Canada. *Water Research, 202*, 117425.

Patrick, R. J., Grant, K., & Bharadwaj, L. (2019). Reclaiming Indigenous Planning as a Pathway to Local Water Security. *11*(5), 936.

Petrea, Ș.-M., Costache, M., Cristea, D., Strungaru, Ș.-A., Simionov, I.-A., Mogodan, A., . . . Cristea, V. (2020). A machine learning approach in analyzing bioaccumulation of heavy metals in turbot tissues. *Molecules, 25*(20), 4696.

Prince George Citizen. (2021). Lheidli T'enneh First Nation calls for federal funding for clean drinking water.

https://www.princegeorgecitizen.com/local-news/lheidli-tenneh-first-nation-calls-for-federal-funding-for-clean-drinking-water-4778801

Qin, Z., Liu, Z., Zhu, P., & Xue, Y. (2020). A GAN-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine, 195*, 105568.

Rasheed, T., Shafi, S., & Sher, F. (2022). Smart nano-architectures as potential sensing tools for detecting heavy metal ions in aqueous matrices. *Trends in Environmental Analytical Chemistry*, e00179.

Revathi, M., Jeya, I. J. S., & Deepa, S. N. (2020). Deep learning-based soft computing model for image classification application. *Soft Computing, 24*(24), 18411-18430.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Rooki, R., Doulati Ardejani, F., Aryafar, A., & Bani Asadi, A. (2011). Prediction of heavy metals in acid mine drainage using artificial neural network from the Shur River of the Sarcheshmeh porphyry copper mine, Southeast Iran. *Environmental Earth Sciences, 64*, 1303-1316.

Rowles III, L. S., Hossain, A. I., Ramirez, I., Durst, N. J., Ward, P. M., Kirisits, M. J., . . . Saleh, N. B. (2020). Seasonal contamination of well-water in flood-prone colonias and other unincorporated US communities. *Science of the Total Environment, 740*, 140111.

Schimpf, C., & Cude, C. (2020). A systematic literature review on water insecurity from an Oregon public health perspective. *International Journal of Environmental Research and Public Health, 17*(3), 1122.

Schwartz, H., Marushka, L., Chan, H. M., Batal, M., Sadik, T., Ing, A., . . . Tikhonov, C. (2021). Metals in the drinking water of First Nations across Canada. *Canadian Journal of Public Health, 112*(Suppl 1), 113-132.

Shafqat, S. S., Rizwan, M., Batool, M., Shafqat, S. R., Mustafa, G., Rasheed, T., & Zafar, M. N. (2023). Metal organic frameworks as promising sensing tools for electrochemical detection of persistent heavy metal ions from water matrices: A concise review. *Chemosphere*, 137920.

Shahi, N. K., Maeng, M., & Dockko, S. (2020). Models for predicting carbonaceous disinfection by-products formation in drinking water treatment plants: a case study of South Korea. *Environmental Science and Pollution Research, 27*, 24594-24603.

Shao, S., Wang, P., & Yan, R. (2019). Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry, 106*, 85-93.

Shen, L., & Qian, Q. (2022). A virtual sample generation algorithm supporting machine learning with a small-sample dataset: A case study for rubber materials. *Computational Materials Science, 211*, 111475.

Shen, Z., Ouyang, X., Xiao, B., Cheng, J.-Z., Shen, D., & Wang, Q. (2023). Image synthesis with disentangled attributes for chest X-ray nodule augmentation and detection. *Medical Image Analysis, 84*, 102708.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1-48.

Singh, R., Gautam, N., Mishra, A., & Gupta, R. (2011). Heavy metals and living systems: An overview. *Indian J Pharmacol, 43*(3), 246-253.

Singh, V., Singh, N., Rai, S. N., Kumar, A., Singh, A. K., Singh, M. P., . . . Mishra, V. (2023). Heavy Metal Contamination in the Aquatic Ecosystem: Toxicity and Its Remediation Using Eco-Friendly Approaches. *Toxics, 11*(2), 147.

Statistics Canada. (2022). Indigenous population continues to grow and is much younger than the non-Indigenous population, although the pace of growth has slowed. https://www150.statcan.gc.ca/n1/daily-quotidien/220921/dq220921a-eng.htm

Stride, B., Abolfathi, S., Odara, M., Bending, G. D., & Pearson, J. (2023). Modelling microplastic and solute transport in vegetated flows Dispersion of polyethylene in submerged model canopies. *Water Resources Research*, e2023WR034653.

Sun, Y., Chen, F., Zafar, A., Khan, Z. I., Ahmad, K., Ch, S. A., . . . Nadeem, M. (2023). Assessment of potential toxicological risk for public health of heavy metal iron in diverse wheat varieties irrigated with various types of waste water in South Asian country. *Agricultural Water Management, 276*, 108044.

Taghizadeh-Mehrjardi, R., Fathizad, H., Ali Hakimzadeh Ardakani, M., Sodaiezadeh, H., Kerry, R., Heung, B., & Scholten, T. (2021). Spatio-temporal analysis of heavy metals in arid soils at the catchment scale using digital soil assessment and a random forest model. *Remote Sensing, 13*(9), 1698.

Teng, Z., Yuan Huang, J., Fujita, K., & Takizawa, S. (2001). Manganese removal by hollow fiber micro-filter. Membrane separation for drinking water. *Desalination, 139*(1), 411-418.

Tremblay, C. V., Beaubien, A., Charles, P., & Nicell, J. A. (1998). Control of biological iron removal from drinking water using oxidation-reduction potential. *Water science and technology, 38*(6), 121-128.

The Council of Canadians. (2020). Guidelines for drinking-water quality: Fourth edition

incorporating the first and second addenda.

https://canadians.org/analysis/fighting-covid-19-starts-universal-access-water-and-

sanitation

The World Bank. (2021). Use of AI Technology to Support Data Collection for Project

Prepa ration and Implementation: A 'Learning-by-doing' Proces.

https://gpss.worldbank.org/sites/gpss/files/knowledge_products/2021/Use%20of%20AI%

20technology%20to%20support%20data%20collection.pdf

Tyagi, S., & Talbar, S. N. (2022). CSE-GAN: A 3D conditional generative adversarial network with concurrent squeeze-and-excitation blocks for lung nodule segmentation. *Computers in Biology and Medicine, 147*, 105781.

Uddin, M. G., Nash, S., Diganta, M. T. M., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management, 321*, 115923.

United Nations. (2015). The human rights to safe drinking water and sanitation : resolution / adopted by the General Assembly. https://digitallibrary.un.org/record/821067

United Nations. (2023). Indigenous Peoples. https://social.desa.un.org/issues/Indigenous-peoples

Valko, M., Morris, H., & Cronin, M. (2005). Metals, toxicity and oxidative stress. *Current medicinal chemistry, 12*(10), 1161-1208.

Veschetti, E., Achene, L., Ferretti, E., Lucentini, L., Citti, G., & Ottaviani, M. (2010). Migration of trace metals in Italian drinking waters from distribution networks. *Toxicological & Environmental Chemistry, 92*(3), 521-535.

Wang, H., Wei, L., Yang, C., Liu, J., & Shen, J. (2020). A pyridine-Fe gel with an ultralow-loading Pt derivative as ORR catalyst in microbial fuel cells with long-term stability and high output voltage. *Bioelectrochemistry, 131*, 107370.

Wang, J., Ji, H., Wang, Q. g., Li, H., Qian, X., Li, F., & Yang, M. (2017). Prediction of size-fractionated airborne particle-bound metals using MLR, BP-ANN and SVM analyses. *Chemosphere, 180*, 513-522.

Wang, J., Yang, Z., Zhang, J., Zhang, Q., & Chien, W.-T. K. (2019). AdaBalGAN: An improved generative adversarial network with imbalanced learning for wafer defective pattern recognition. *IEEE Transactions on Semiconductor Manufacturing, 32*(3), 310-319.

Wang, R., Kim, J.-H., & Li, M.-H. (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Science of the Total Environment, 761*, 144057.

Winter, E. (2002). The shapley value. *Handbook of game theory with economic applications, 3*, 2025-2054.

Wolfe, P. (2006). Settler Colonialism and the Elimination of the Native. *Journal of genocide research, 8*(4), 387-409.

Woodcock, G. (1988). A social history of Canada. Toronto: Viking Penguin Group.

World Health Organization. (2017). Guidelines for drinking-water quality: fourth edition incorporating first addendum, 4th ed + 1st add. https://iris.who.int/handle/10665/254637

World Health Organization. (2023). Drinking-water. https://www.who.int/news-room/fact-sheets/detail/drinking-water

Yan, X., Zang, Z., Luo, N., Jiang, Y., & Li, Z. (2020). New interpretable deep learning model to monitor real-time PM2. 5 concentrations from satellite data. *Environment International, 144*, 106060.

Yeganeh, M., Azari, A., Sobhi, H. R., Farzadkia, M., Esrafili, A., & Gholami, M. (2023). A comprehensive systematic review and meta-analysis on the extraction of pesticide by various solid phase-based separation methods: a case study of malathion. *International Journal of Environmental Analytical Chemistry, 103*(5), 1068-1085.

Zhai, Y., Han, Y., Xia, X., Li, X., Lu, H., Teng, Y., & Wang, J. (2021). Anthropogenic Organic Pollutants in Groundwater Increase Releases of Fe and Mn from Aquifer Sediments: Impacts of Pollution Degree, Mineral Content, and pH. *Water, 13*(14), 1920.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, H., Yin, S., Chen, Y., Shao, S., Wu, J., Fan, M., . . . Gao, C. (2020). Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. *Journal of cleaner production, 273*, 122858.

Zhang, P., Yang, M., Lan, J., Huang, Y., Zhang, J., Huang, S., . . . Ru, J. (2023). Water Quality Degradation Due to Heavy Metal Contamination: Health Impacts and Eco-Friendly Approaches for Heavy Metal Remediation. *Toxics, 11*(10), 828.

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST), 8*(3), 1-19.

Zhang, Y., Wang, Z., Zhang, Z., Liu, J., Feng, Y., Wee, L., . . . Traverso, A. (2023). GAN-based one dimensional medical data augmentation. *Soft Computing*, 1-11.

Zhao, B., & Yuan, Q. (2021). Improved generative adversarial network for vibration-based fault diagnosis with imbalanced data. Measurement, 169, 108522.

Zhao, J., Yan, X., Zhu, T., Wang, J., Li, H., Zhang, P., . . . Ding, L. (2015). Multi-throughput dynamic microwave-assisted leaching coupled with inductively coupled plasma atomic emission spectrometry for heavy metal analysis in soil. *Journal of Analytical Atomic Spectrometry, 30*(9), 1920-1926.

Zoni, S., & Lucchini, R. G. (2013). Manganese exposure: cognitive, motor and behavioral effects on children: a review of recent findings. *Current opinion in pediatrics, 25*(2), 255.

Assembly of First Nations. (2021). Description of the AFN. https://www.afn.ca/description-of-the-afn/