EXPLORATORY ANALYSIS OF WATER QUALITY IN A SMALL, URBANIZED WATERSHED USING DEEP LEARNING

ALFRED OFOSU

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND
TECHNOLOGIES

YORK UNIVERSITY
TORONTO, ONTARIO

August 2023

Abstract

Water is a life-sustaining resource for living organisms inside and outside water bodies. Natural waters serve as municipal and industrial water supplies, sources for agricultural irrigation, homes for aquatic ecosystems, recreation, and other essential uses. The quality of water determines its use. Therefore, it must be monitored, managed, and reported to help stakeholders in decision-making that can protect watershed ecosystems and improve measures to mitigate factors adversely affecting water bodies. Water quality is represented by a set of parameters that describe specific characteristics or properties of water. These parameters are determined by measuring water's physical and chemical characteristics and concentration levels of various substances in a water column with subsequent sample analysis in laboratories. This results in low frequencies of observations for water quality parameters compared to hydrometric and meteorological data. Frequencies of observation adopted by many water quality monitoring systems vary between 4 and 12 samples per year, suggesting applying modelling techniques to support decision-making.

The study aims to develop a data-driven computational tool for water quality modelling in a small, highly urbanized watershed of the Don River, Ontario, Canada. The study focuses on major ions, namely, cations: calcium (Ca2+), magnesium (Mg2+), sodium (Na+), and potassium (K+), and anions such as bicarbonate (HCO3-), carbonate (CO32-), chloride (Cl-), and sulphate (SO42-). These parameters are not affected significantly by the aquatic ecosystem. The hydrological and meteorological processes mainly determine their dynamics. The study uses data from different monitoring systems belonging to the Toronto and Region Conservation Authority (TRCA) and Environment and Climate Change Canada (ECCC). It consists of water quality parameters and hydrometric and meteorological characteristics observed in the watershed over 57 years. Concentrations of

selected water quality parameters are modelled using deep neural networks. The data pre-processing framework for cleansing and integrating data observed at different frequencies from different locations is developed. The framework is applied for the comparative analysis of neural networks of various configurations.

Two sets of computational experiments were conducted. In the first set of experiments, integrated data from all monitoring stations in the watershed was fed into the deep learning algorithms to train a neural network to predict the concentration of major ions for the upcoming month (t+1). The second set of experiments uses upstream environmental parameters to train the model and predict the major ion concentrations in the lower subwatershed. The study investigates the performance of developed models in accurately predicting ion concentrations and provides insights into the relationship between environmental factors and water quality in the investigated watershed. The findings have practical applications for water resource management and pollution prevention efforts.

Keywords: *Water quality, integrated hydrology, major ions, deep learning*

Dedication

To my daughters, Avielle and Alayjah. I hope they read it someday.

Acknowledgement

I want to express my sincere gratitude to Prof. Marina Erechtchoukova for her invaluable guidance, support, and mentorship throughout the entire process of completing this thesis. Your expertise, patience, and insightful feedback have been instrumental in shaping my research and refining my understanding of machine and deep learning. Your dedication to fostering intellectual growth and your unwavering encouragement have been truly inspiring. Thank you for being a source of inspiration and guiding me toward achieving excellence in my academic pursuits.

I want to thank Prof. Augustine Chi Mou Wong and Prof. Adeyemi Oludapo Olusola for their valuable contributions as committee members in evaluating my thesis. Your expertise, critical insights, and constructive feedback have played a crucial role in enhancing the quality of my research work.

I want to express my deepest gratitude to my parents, Mr. and Mrs. Poku, whose unwavering love, sacrifices, and encouragement have been the foundation of my journey. Your constant support and belief in my abilities have been my guiding light. I want to express my heartfelt appreciation to my Aunt, Dr. J.O. Dankyi, Uncle, Mr. K. Dankyi, and Aunt, Mrs. S. Kuffour-Berko, for their unwavering encouragement that has always inspired me to pursue excellence. To Grandma B. Forster, your prayers and guidance have been a guiding light throughout my childhood. I sincerely thank my brother, F. F. Ofosu, for his continuous support and prayers. I am also profoundly thankful to Uncle S. Atinkah and Aunt G. Atinkah for their constant encouragement and support throughout my academic journey. To everyone who has played a part in my journey, I extend my heartfelt thanks for your unwavering support and belief in me.

To my loving wife, Evelyn, your patience, understanding, and encouragement have been my source of strength. Your unwavering support and sacrifices have enabled me to pursue my academic goals diligently. Thank you for always standing by my side.

To my two wonderful daughters, Avielle and Alayjah, you fill my life with joy, laughter, and inspiration. Your presence reminds me of the importance of balance and to cherish every moment. Your smiles are my motivation. I am truly blessed to have a family that believes in me, supports me, and celebrates my successes. Your love has been my driving force, and I am deeply grateful for the sacrifices you've made to help me achieve my dreams. Thank you for being my pillars of strength.

Table of Contents

List of Tables

List of Figures

1 Introduction

Water is a fundamental resource for supporting life within and outside water bodies. Natural water sources play critical roles as suppliers of municipal and industrial waters, irrigation for agriculture, habitats for diverse aquatic ecosystems, and venues for recreation and various vital activities. The suitability of water for multiple uses is directly linked to its quality, making water quality assessment, management, and reporting crucial tasks to support informed decision-making.

Water quality in natural sources is determined by a collection of parameters encompassing specific water characteristics and properties. These parameters represent water's biological, physical, and chemical attributes and are derived through observations and measurements. Many water quality parameters reflect the concentration of various substances in a water column. The quality of natural waters is observed and reported by specialized monitoring systems comprising monitoring sites established at different locations on natural streams and reservoirs. On-site automated devices measure some water quality parameters, while most other parameters can be determined only through lab analytical procedures (Ofosu & Erechtchoukova, 2023). Water samples are collected routinely from different monitoring sites and subjected to laboratory analysis to obtain comprehensive information about their condition.

Unlike hydrometric and meteorological data, which are recorded more frequently, many water quality parameters often have lower observation frequencies, ranging from 4 to 12 samples per year. Some monitoring authorities collect samples during the ice-free periods (April to November) and just a few additional samples during winter (SSEA, 2002). To compensate for such sparse data, employing modelling techniques becomes crucial to support decision-making processes effectively.

Additionally, the quality of water in natural waterbodies is influenced by various factors, including the hydrological characteristics of its watershed, prevailing weather conditions, the health of the aquatic ecosystem, human activities (anthropogenic impact), and land use within the watershed. Traditional process-based models attempt to capture these factors through complex mathematical expressions to assess and model water quality. However, these models heavily rely on numerous parameters unique to each watershed and waterbody, requiring extensive field observations and measurements of morphometric characteristics, land cover, and land uses.

In contrast, data-driven approaches offer a more efficient alternative by utilizing routinely collected data from various monitoring systems, provided rich datasets are available. This study explores the feasibility of applying deep learning to modelling water quality in natural waters solely using data provided by routine monitoring systems. This approach avoids the time-consuming and costly process of gathering specific watershed parameters, making it an attractive option for assessing water quality dynamics.

Through applying deep learning techniques, this research seeks to uncover valuable insights into the potential of data-driven approaches for predicting and understanding water quality variations in natural water bodies where water samples are collected monthly. To mitigate data scarcity, this study integrates hydrometric and meteorological data with water quality data to (1) impute missing values and (2) explore the possibility of improving the predictive ability of constructed models. By leveraging state-of-the-art modelling approaches, the research aims to establish a comprehensive framework that enhances water quality management and decision-making processes despite the high sparsity and intermittencies in the data.

The study focuses on modelling water quality of the Don River in Ontario, Canada and specifically, concentrations of major ions such as cations: calcium (Ca^{2+}), magnesium

(Mg2+), sodium (Na+), and potassium (K+), and anions: carbonate (CO32-), chloride (Cl-), and sulphate (SO42-). These constituents contribute to the amount of Total Dissolved Solids (TDS) and specific conductance in waters (Bartos & Ogle, 2002). Their conservative nature drives the choice of these water quality parameters in a water column. The aquatic ecosystem does not significantly influence these parameters, and hydrological and meteorological processes and anthropogenic impacts primarily govern their dynamics (Ofosu & Erechtchoukova, 2023).

The Don River is a small river with a highly urbanized watershed. A fast response to water and pollution load characterizes the stream. This makes modelling the hydrochemical regime of this river very challenging.

The study uses data collected from different monitoring systems: (1) the water quality monitoring system belonging to the Toronto and Region Conservation Authority (TRCA), (2) the Provincial Water Quality Monitoring Network (PWQMN); the meteorological system belonging to the (3) Meteorological Service Canada (MSC), and the hydrometric system belonging to the (4) Water Survey Canada (WSC).

A data preprocessing framework for cleansing, transforming, and integrating data with different observation frequencies is developed and applied to different deep learning model configurations for comparative analysis.

The research conducted two sets of computational experiments to investigate water quality prediction in the watershed. In the first set of experiments, data from all monitoring stations in the watershed were integrated and utilized to train deep learning algorithms. The objective was to develop a neural network capable of predicting the concentration of a specific major ion for the upcoming month (t+1). In the second set of experiments, the focus shifted to using upstream parameters as inputs to the model to predict major ion concentrations in the lower basin of the watershed.

The study rigorously evaluated various modelling techniques, focusing on the potential of deep learning techniques to enhance the models' predictive capabilities. Deep learning algorithms were employed to explore their effectiveness in improving the accuracy and performance of the models. Through this comprehensive evaluation, the research aimed to identify the most suitable and effective approaches for predicting water quality parameters, ultimately contributing to the advancement of predictive modelling in water quality research.

The thesis is organized in the following way: Chapter 2 explores the different concepts of deep learning, their architectures, their current state, and their potential in modelling water quality with sparse data. Chapter 3 presents a comprehensive framework that describes the steps from data acquisition to preprocessing, modelling, and predicting the outcomes of water quality parameters with data from various sources. It outlines the steps for preparing data with different observation frequencies. Furthermore, a detailed explanation of selecting and preparing features for deep learning models is discussed. In Chapter 4, the computational experiments, the comparative analysis, and the obtained results are presented. Finally, Chapter 5 presents the conclusions drawn based on the study's outcomes and potential directions for future research.

2 Theoretical Background, Related work, and Research Questions

2.1 Deep Learning Concepts

The term "Artificial Intelligence" was introduced in the 1950s by a small group of pioneers in the emerging field of computer science. These early visionaries began pondering the possibility of creating computers exhibiting human-like cognitive abilities, sparking the exploration of whether machines could be designed to "think." For a considerable period, this notion of a computer having human-like abilities incited a belief among experts that achieving artificial intelligence at a human level could be accomplished by manually crafting a substantial set of explicit rules for manipulating knowledge stored in special databases. This approach, known as symbolic AI, dominated the field of artificial intelligence from the 1950s until the late 1980s. Symbolic AI reached its pinnacle of popularity during the expert systems boom of the 1980s (Chollet, 2022).

In a nutshell, AI is the development of intelligent systems that can perform tasks that typically require human intelligence. It encompasses various techniques, approaches, and technologies that aim to simulate human intelligence in machines by perceiving and understanding their environment, reasoning and making decisions and learning and adapting from experiences. AI also encompasses other areas, such as natural language processing, computer vision, expert systems, and robotics, as seen in Figure 1.

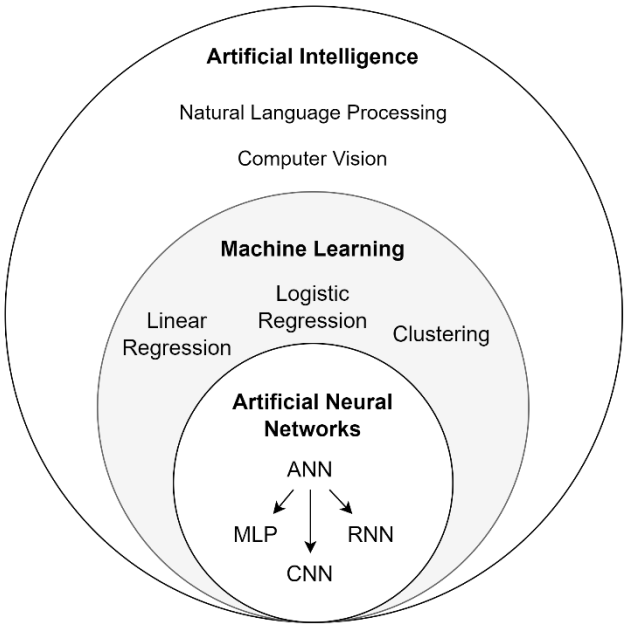


Figure 1 Artificial Intelligence, Machine Learning, and Artificial Neural Networks

Machine Learning (ML) is a key component of AI. The term "learning" in the context of machine learning refers to the automatic mapping of independent variables to a dependent variable based on a comprehensive dataset (Chollet, 2022). In other words, a machine learning algorithm can analyze and process data to identify patterns, relationships, and trends between the input variables (independent variables) and the output variable (dependent variable).

Given a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ that maps a n-dimensional set of real numbers to a real number the relationship between input variables and to a target can be denoted as:

$$f(x) = \lambda(\alpha \cdot x + \beta) \qquad \textit{Equation 1}$$

where x is the input variable, $\lambda : \mathbb{R} \rightarrow \mathbb{R}$ is a continuous non-linear function called the activation function, $\alpha \in \mathbb{R}^n$ is a vector of weights and scaler $\beta \in \mathbb{R}$ is called the bias. Here $\alpha \cdot x$ is the inner product on $\mathbb{R}^n$. $f(x)$ is often denoted as y, the output variable. By observing a wide range of samples i.e., the mappings of input values to output values, as shown in Figure 2 , the algorithm can learn the underlying mapping function and make predictions or decisions on unseen data.

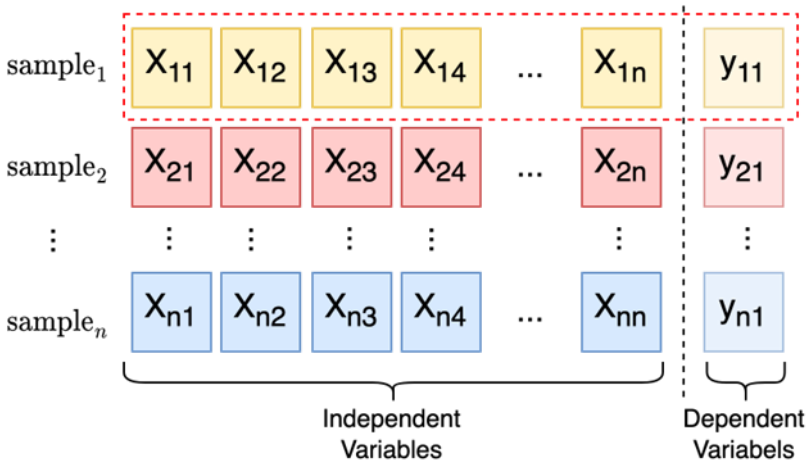


Figure 2 Mapping of input values to output values.

The learning process involves extracting meaningful insights from the data and generalizing from the observed samples to accurately predict the output variable for new, unseen instances.

Machine learning can be broadly categorized into three main branches: Supervised Learning, Unsupervised Learning and Reinforcement Learning. In supervised learning, the algorithm is trained on labelled data where each data instance is associated with a corresponding target or output value. The algorithm learns to map the input features (independent variables) to the desired output (dependent variables) based on the labelled examples. The objective is to generalize from the training data and accurately predict the output for new, unseen data. Classification and regression are common tasks performed in supervised learning.

Unsupervised learning involves training algorithms on unlabeled data, where the input features are provided without any corresponding output labels. The algorithm learns to identify patterns, structures, and relationships in the data without specific guidance. Clustering, dimensionality reduction, and anomaly detection are common tasks in unsupervised learning.

Reinforcement learning involves an agent learning to make sequential decisions in an environment to maximize a reward signal. The agent interacts with the environment, takes actions, and receives feedback through rewards or penalties. Through trial and error, the agent learns to optimize its decision-making strategy to achieve the maximum cumulative reward over time.

These branches of machine learning provide different approaches and techniques to tackle various problems, allowing for the development of intelligent systems that can learn from data and make informed decisions.

Two main algorithms are commonly used in the supervised learning paradigm: classification and regression. The key distinction between these two algorithms lies in the nature of the dependent variable they handle. Classification algorithms are used when the dependent variable is categorical or discrete. The goal is to assign input data points to specific predefined classes or categories based on the features or attributes. The algorithm learns from labelled data, where each data instance is associated with a class label. Examples of classification algorithms include logistic regression, decision trees, support vector machines (SVM), random forests and neural networks.

Regression algorithms, on the other hand, are employed when the dependent variable is continuous or numerical. The objective is to predict a numerical value or estimate a relationship between the input and output variables. Linear regression, polynomial regression, support vector regression (SVR), and neural networks are common examples of regression algorithms. While both classification and regression algorithms fall under the supervised learning umbrella, they cater to different types of dependent variables.

In the context of this research, this thesis focuses on the prediction of continuous variables, which categorizes the task as a regression problem. Consequently, all diagrams and tables presented within this research will reflect this regression context.

Deep Learning (DL), a specific branch of supervised machine learning, started gaining significant attention and popularity in the data science community in 2010. This is because ML has its limitations when it comes to perceptual tasks such as vision, sight, and language (Chollet, 2022). Deep learning takes the principles of machine learning further by introducing a network architecture – an Artificial Neural Network (ANN) – composed of shallow or deep interconnected layers, with many nodes or units within each layer. DL algorithms are designed to automatically learn and extract intricate patterns, representations, and features from complex datasets. The architecture of deep

learning models typically consists of an input layer, multiple hidden layers, and an output layer, see Figure 4. The hidden layers in deep learning networks enable the extraction of hierarchical representations, allowing the model to capture low-level and high-level features from the data.

DL has achieved remarkable success in various domains, surpassing traditional machine learning methods in accuracy and performance on challenging tasks. Deep learning models have achieved breakthrough results in image classification, object detection, machine translation, and many other domains.

On the downside, deep learning models require significant computational resources and large amounts of labelled training data. Training deep learning models can be computationally intensive and may necessitate specialized hardware. Additionally, interpreting and understanding the internal operations of deep learning models can be challenging due to their complex nature and black-box-like behaviour. Nonetheless, deep learning continues to advance the field of machine learning. It has revolutionized various industries by enabling the development of highly sophisticated and intelligent systems such as search engines like Google, which autonomously suggest search queries and automated chat system like ChatGPT that interacts with humans using natural language.

Understanding the behaviour of a single artificial neuron, or perceptron, is foundational to comprehending a neural network. It involves the computation of weighted inputs, applying an activation function, and adjusting weights during training. A perceptron represents a simplified model of a biological neuron, Figure 3, which takes multiple input values $(x_1, x_2, \cdots, x_n)$ each multiplied by its corresponding weight, $(\alpha_1, \alpha_2, \cdots, \alpha_n)$, a bias, β, and calculates a weighted sum of these inputs. It then applies an activation function to the weighted sum to produce an output (*Equation 1*). Then, the output is passed to the next layer of neurons or used as the final prediction of the perceptron. In Figure 4, the output

layer contains just one neuron, which outputs a continuous value. The weights associated with the inputs determine the strength and importance of each input in the computation. During the learning process, these weights are adjusted to optimize the performance of the perceptron. The adjustment of weights is performed using a technique called backpropagation, where the error between the predicted output and the desired output determines the rate of change in the weights. Suppose the goal is to accept the differences AS-IS after each sample is analyzed without optimizing the weights and biases to improve the model's accuracy. In that case, we have what is called a feed-forward neural network.

An activation function, sometimes referred to as a transfer function, determines the activation of a neuron and adds nonlinearity features to a neural network; this overcomes the restrictions of conventional linear regressions and allows the network to capture complex patterns within data and through that, the expressive capability of the network is increased, resulting in more accurate representation (Adu et al., 2022).

Common activation functions include the Sigmoid, Hyperbolic Tangent (tanh), and Rectified Linear Unit (ReLU) functions. An ANN is formed by connecting multiple perceptrons in a layered structure. Each perceptron in a layer receives inputs from the previous layer and produces outputs that serve as inputs for the next layer. This interconnectedness enables ANNs to learn and predict complex patterns and relationships within data.

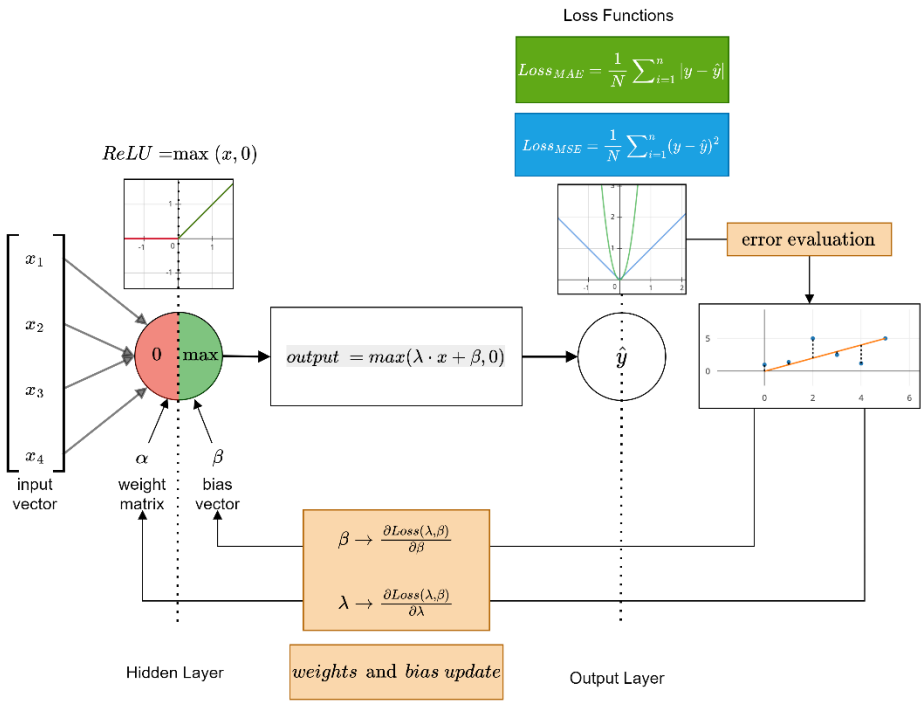


Figure 3 A deep learning algorithm explained with a perceptron.

A loss function called a cost or objective function, calculates the difference between the predicted target value and the actual target value for each training sample. If the error or variance is substantial, the optimization algorithm performs partial differentiation of the loss function with respect to the weights and biases. The primary objective is minimizing the loss, enhancing the model's predictive accuracy and reducing prediction errors (Nielsen, 2015). Common optimization algorithms include Stochastic Gradient Descent (SGD), Root Mean Square Propagation (RMSProp), and Adaptive Momentum Estimation (Adam). This process of updating the weight and bias continues until a local or global minimum is reached during the partial differentiation of the loss function.

There are several ANN architectures with unique features, suited best for applications (Simulink, 2022). Some of the common ANN architectures include Deep Neural Network (DNN), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

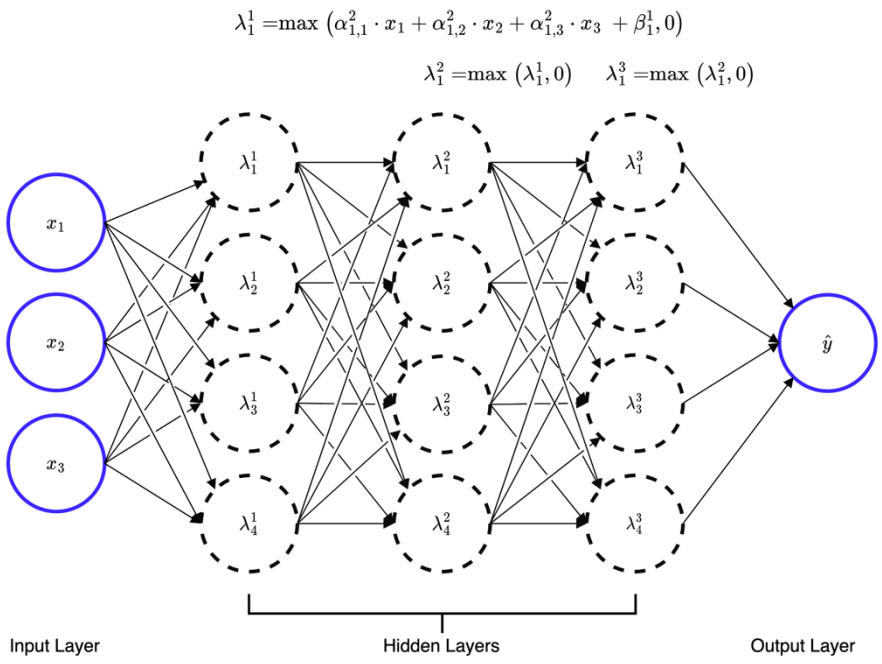


Figure 4 A densely connected neural network architecture with hidden layers.

When an ANN has more than one hidden layer, it is called a DNN. A DNN architecture is depicted in Figure 4. The depth of the hidden layers is what is referred to as "Deep" in the context of DL.

In a typical DNN, each neuron in the input layer is connected to a neuron in the hidden layer; however, in the CNN architecture, only a small region, also called local receptive fields, of input layer neurons connect to the hidden layer. CNNs were developed for computer vision computations, such as classifying images. In the case of computer vision, the local receptive field is translated across an image (a 2D image); in other words, it convolves over an image to create a feature map from the input layer to the hidden layer. With respect to a one-dimensional input, such as in the case of this research, the CNN will translate across a sequence of input values to create a feature map from the input layer to the hidden layer.

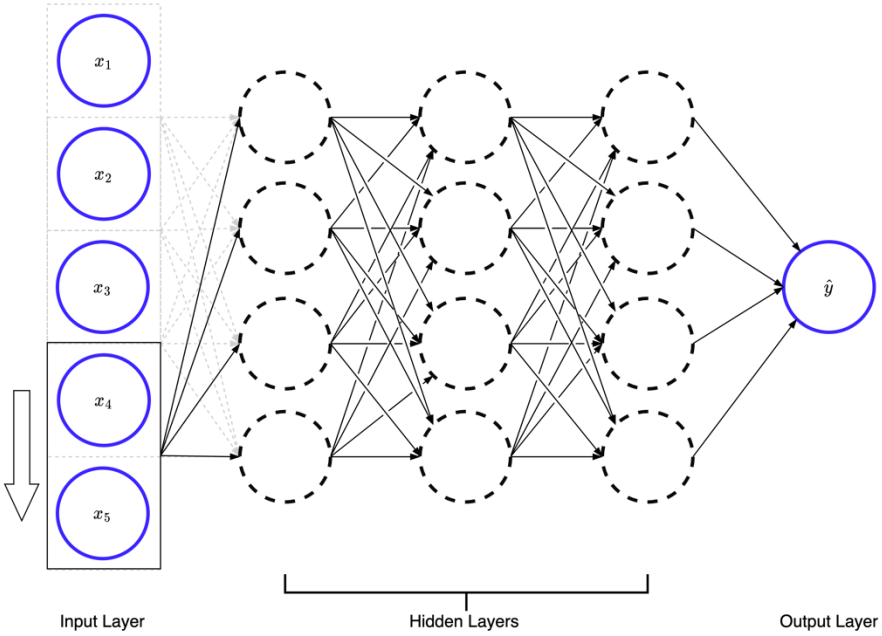


Figure 5 A one-dimensional convolutional neural network architecture.

In contrast to DNNs, the shared weights and biases in the CNN hidden layers are the same in each layer, meaning each layer detects the same feature. An activation function is applied to each convolutional hidden layer to generate output values over the entire set of input values. A further step called pooling is used to reduce the dimensionality of the convolutional output values by condensing the output of small regions of neurons into a single output. This helps to reduce the number of parameters sent over to the fully connected layer or output neuron (Figure 5).

RNNs are built differently than DNNs and CNNs because they can store memory as they process sequences of data inputs. It maintains a state (γ) that contains information relative to what it has seen so far and improves the performance of the network on current and future inputs (Figure 6). In a nutshell, RNNs are networks with loops, allowing information to persist (Olah, 2015).

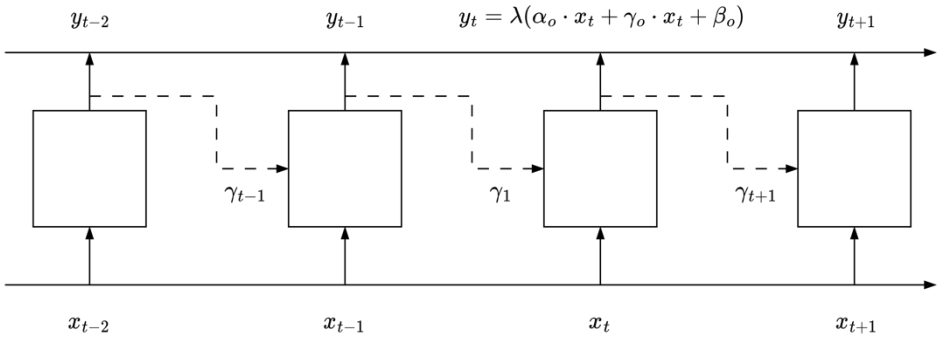


Figure 6 A recurrent neural network architecture.

Within the broader category of RNN architectures, there are several subtypes designed to address specific challenges associated with processing sequential data. Common types are Simple RNN (Figure 6), Gated Recurrent Unit (GRU) (Cho et al., 2014), and Long Short-Term Memory (LSTM) unit (Hochreiter & Schmidhuber, 1997). Simple RNN suffers under the vanishing gradient problem and hence cannot capture long-term dependencies in sequential data. LSTM was designed with three gating mechanisms: the forget gate, input gate, and output gate. These gates allow the model to decide what information to forget, what new information to add, and what information to output. This architecture is explicitly designed to capture and retain information over long sequences, making it especially effective for tasks involving complex dependencies across time steps. With that in mind, it made LSTMs particularly relevant for modeling the water quality parameter.

2.2 Literature Review

Shelton et al., 2021 employed three machine learning algorithms - Random Forest, k-Nearest Neighbors, and Naïve Bayes - to investigate the feasibility of accurately classifying a produced water sample to a specific geologic watershed based on similarities to a provided training dataset. While they achieved reasonably high accuracies of around 75%, their approach had certain limitations. Notably, they did not consider the time-relatedness of the water quality parameters, potentially overlooking important temporal

patterns in the data. Additionally, they chose to drop parameters with missing values, which could introduce uncertainty into the model's predictions. Omitting certain parameters from the analysis might adversely affect the model's performance and could hinder a comprehensive understanding of the water quality patterns in the investigated system.

Boccadoro et al., 2022 utilized the Long Short-Term Memory (LSTM) model to predict water quality parameters such as pH, conductivity, oxygen, and temperature. The LSTM model demonstrated exceptional performance, achieving high accuracy and low errors with a Mean Absolute Error of 0.20, a Mean Square Error of 0.092, and a Cosine Proximity of 0.94. However, the researchers encountered a significant sparsity level in their data, leading them to exclude some years from their analysis. Additionally, they selected only three readings per day at specific times (9 am, 12 pm, and 6 pm) for their research. This approach may result in uneven time frequencies, potentially limiting the neural network's ability to represent the data's temporal patterns comprehensively. Also, excluding some water quality parameters from a model may hinder a model's ability to capture latent representations in the data.

Duhayyim et al., 2022 developed a new smart water quality prediction using atom search optimization with the fuzzy deep convolution network (WQP-ASOFDCN) technique in the IoT environment, which achieved an accuracy of 98% and proved to be better than machine learning algorithms such as Light Gradient Boosting Machine (LGBM), eXtreme Gradient Boosting (XGBoost), Naïve Bayes, K-Nearest Neighbor (KNN) and Random Forest that were compared. However, there is no indication on the observation frequencies of how the water quality parameters, i.e., pH, temperature, and turbidity, were captured and, thus, neglected time series relatedness in the dataset, which may hinder an ANN from understanding temporal patterns. Additionally, missing values

were filled with mean values and no other experiments on other state-of-the-art mechanisms designed to handle missing data were investigated.

Yao et al., 2022 explored the prediction of the long-term integrated water quality index in the Chaohu Lake area using various popular deep learning models, such as Multi-layer Perceptrons (MLP), Long Short-Term Memory (LSTM), and Transformer-based models. The experimental findings showed that all the chosen models performed well and demonstrated promising predictive capabilities within the study area. Notably, the Transformer-based model exhibited superior performance compared to the other models. Despite the good results, certain aspects of the study needed more clarity. One notable concern was the need for a clearer indication of how the data was preprocessed, especially considering the daily observation frequency of the data selected from 2019 to 2022. Transparent and detailed data preprocessing steps are essential for understanding the reliability and robustness of the model's outcomes. Additionally, the study needed to provide explicit information on how the nine water quality parameters were measured and tested to ensure the accuracy and reliability of the data used in the analysis. A thorough understanding of the data quality and measurement procedures is critical for drawing valid conclusions and making informed decisions based on the study's findings. Providing such information would enhance the credibility and reproducibility of the research results.

Aldhyani et al., 2020 employed advanced artificial neural network models, specifically the nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM), along with machine learning algorithms such as support vector machine (SVM), K-nearest neighbour (KNN), and Naive Bayes. These models were used to predict water quality index (WQI) and water quality classification (WQC), respectively, and both neural networks achieved impressive accuracies, exceeding 94%. However, despite the

promising results, the research needed to thoroughly investigate the time series aspect of the water quality data. The temporal patterns that could be learned by the models about the variations in water quality indices overtime needed to be given more attention. Neglecting this crucial temporal aspect may limit the models' ability to capture and utilize time-dependent trends and patterns in the water quality data, potentially leading to missed opportunities for more accurate predictions and insights. Addressing the time series aspect of the data could enhance the models' predictive performance and provide valuable information about how water quality varies over time. It is essential to consider the temporal relationships between different water quality measurements and their impact on the overall water quality index. A comprehensive analysis of the temporal patterns would contribute to a more robust and accurate prediction of water quality and classification, enabling more informed decision-making and better water resource management strategies.

Wu et al., 2022 propose a water quality prediction model utilizing multi-task deep learning, specifically focusing on the chemical oxygen demand (COD) in the water environment of the Lanzhou region of the Yellow River. Their model adopts a novel approach by simultaneously training and learning multiple sections (monitoring stations) of correlation, allowing for sharing of water quality information across different areas while preserving their heterogeneity. To extract features from local to full-time series water quality data, they employ a hybrid model called CNN-LSTM, which enhances the mining capabilities of the model. The experimental results show notable improvements when comparing their model's performance of the current single-section (monitoring station) water quality prediction. The predicted values' mean absolute error (MSE) and root mean square error (RMSE) decreased by 13.2% and 15.5%, respectively. However, some limitations in their research warrant consideration. Firstly, the

observation frequency used for monitoring water quality parameters was monthly from 2018 to 2020, which resulted in only 36 months of data. This relatively small dataset may only partially exploit the potential of the employed deep learning architecture. A larger dataset could offer more robust training and validation for the model, potentially leading to even more accurate predictions. Secondly, they split the data into training, validation, and test sets in a 28:3:3 ratio. Such a small proportion of the data used for validation and testing could influence the model's performance and lead to overfitting or biased results. Considering the small dataset, a time series cross-validation approach might have been more appropriate for a more reliable model evaluation. Additionally, the researchers focused only on a small portion of the watershed system and should have considered modelling the entire watershed. This could lead to missing information from other watershed sections, which may be relevant for improving the model's predictive power. Considering the entire watershed system could provide a more comprehensive understanding of the water quality dynamics and lead to more accurate predictions.

Zhou et al., 2022 propose a novel approach for modelling time series water quality data by decomposing it into two key characteristics: trend and fluctuation. To achieve this, they utilize an ensemble method that effectively models the trends and fluctuations separately. The trends are handled using a traditional time series method, the Autoregressive Integrated Moving Average (ARIMA), while the fluctuations are addressed with a recurrent neural network, the Gated Recurrent Unit (GRU). This ensemble model, the W-ARIMA-GRU model, demonstrates superior prediction accuracy, stability, and robustness for three conventional water quality indicators. Moreover, this paper employs the ensemble learning model LightGBM for predicting the water quality evaluation level. The results show remarkable accuracy and F1-score, reaching 97.5% and 97.8%, respectively, indicating strong performance. Despite these high F1-score, deep

learning models could eliminate the need to decompose and remove trends before model training. The research uses daily monitored water quality parameters, but many of these data points need to include values, leading to their removal from the analysis. Omitting certain parameters from the study might have adverse effects on the model's performance and could hinder a comprehensive understanding of the water quality patterns in the investigated system, and addressing missing values more effectively could potentially enhance the model's overall performance and allow for a more accurate and detailed analysis of water quality trends and fluctuations.

2.3 Water Quality of Natural Waters

A watershed, or a drainage or catchment area, refers to the geographic location of land that collects and drains water, including rainfall, snowmelt, and runoff, into a common outlet such as a river, stream, lake, wetland, or estuary. It is defined by the land's topography, with high points (such as hills or mountains) forming the boundaries that separate one watershed from another.

The watershed concept underscores the intricate connection between the natural landscape and the hydrological system. When precipitation occurs, such as rain or snow, the water that descends onto the land within a watershed follows diverse pathways. Some infiltrate the soil, some evaporate, and some flow over the surface. Eventually, this water converges to a river or stream, serving as the outlet for the entire drainage area. Consequently, the quality of water in the stream is influenced by various biological, chemical, and physical components; thus, *water quality* can be defined as the physical, chemical, and biological characteristics of water (Alley, 2007; Spellman, 2008) and can be grouped into three main categories: biological, chemical, and physical. These groups of parameters provide valuable information about the condition and characteristics of a water body, helping to assess its suitability for various uses and to identify potential environmental concerns.

Physical water quality parameters relate to the physical properties of water and provide information about its appearance, clarity, and physical composition. Some commonly measured physical parameters include turbidity, temperature, colour, and sensory characteristics such as taste and smell contaminants (Eaton et al., 2005).

Chemical water quality parameters involve the measurement of various chemical constituents present in the water, providing insights into its chemical composition and

potential contaminants. Some commonly measured chemical parameters include pH, acidity and alkalinity, hardness, chlorine, and dissolved oxygen (Eaton et al., 2005).

Biological water quality parameters focus on the presence and abundance of living organisms and microorganisms, providing insights into the ecological health and potential risks associated with water bodies. Some common biological parameters include bacteria, algae, and viruses (Eaton et al., 2005).

The study focuses on modelling water quality of the Don River in Ontario, Canada, and specifically, concentrations of major ions such as cations: calcium (Ca2+), magnesium (Mg2+), sodium (Na+), and potassium (K+), and anions: carbonate (CO32-), chloride (Cl-), and sulphate (SO42-). These constituents contribute to the amount of Total Dissolved Solids (TDS) and specific conductance in waters (Bartos & Ogle, 2002).

By measuring the concentrations of these ions in surface water samples, the ionic composition of the water is determined, and the chemical quality of the water can be characterized and described. According to (Bartos & Ogle, 2002), the significance of monitoring these elements is listed in the table below.

Table 1 Source or cause, and significance of dissolved-mineral constituents and physical properties of water (Bartos & Ogle, 2002; Popkin, 1973)

Major Ion	**Source or cause**	**Significance**
Calcium (Ca) and Magnesium (Mg)	Dissolved from many rocks and soil, but especially from limestone, dolomite, and gypsum. Calcium and magnesium are detected in large quantities in some brines. Magnesium is present	Large concentrations, in combination with chloride, give a salty taste. Moderate concentrations have little effect on the usefulness of water for most purposes. Sodium salts may cause

	in large quantities in seawater.	foaming in steam boilers. A large sodium concentration may limit the use of water for irrigation.
Sodium (Na) and potassium (K)	Dissolved from many rocks and soil; also in ancient brines, seawater, industrial brines, and sewage.	Large concentrations, in combination with chloride, give a salty taste. Moderate concentrations have little effect on the usefulness of water for most purposes. Sodium salts may cause foaming in steam boilers. A large sodium concentration may limit the use of water for irrigation.
Bicarbonate (HCO3) and carbonate (CO3)	Action of carbon dioxide in water on carbonate rocks such as limestone and dolomite.	Bicarbonate and carbonate produce alkalinity. Bicarbonates of calcium and magnesium decompose in steam boilers and hot-water facilities to form scale and release corrosive carbon dioxide gas. In combination with calcium and magnesium, causes carbonate hardness.

Sulfate (SO4)	Dissolved from rocks and soil containing gypsum, iron sulfides, and other sulfur compounds. Commonly present in mine water and in some industrial wastes.	Sulfate in water containing calcium forms a hard scale in steam boilers. In large concentrations, sulfate in combination with other ions gives bitter taste to water and may have a laxative effect on some people. Some calcium sulfate is considered beneficial in the brewing process.
Chloride (Cl)	Dissolved from rocks and soil. Present in sewage and found in large concentrations in ancient brines, seawater, and industrial brines.	In large concentrations in combination with sodium, gives salty taste to drinking water. In large concentrations increases the corrosiveness of water towards some metals.
Total Dissolved solids	Primarily mineral constituents dissolved from rocks and soil.	Water containing more than 1,000 mg/L dissolved solids is unsuitable for many purposes.

Additionally, (Kumar et al., 2022) emphasized that the presence of these ions in river water can have detrimental effects in the long term, leading to potential harm to fishes, invertebrates, and aquatic plants.

Major ions play a significant role in determining the specific conductance or electrical conductance (CNDT) of water, as highlighted by the U.S. Geological Survey (USGS) in their study on Major Ions (USGS, 2023). (Zhang et al., 2017) revealed in their research that TDS positively impacts EC; thus, it was included in the model to support the predictions of the major ions. Other physical parameters such as Alkalinity (ALKT), Hardness (HARD), pH, and Water Temperature ($temp_w$) were included in the modelling of the major ions.

The inclusion of these specific physical parameters in this research was influenced by the following: Their substantial presence of data values within the dataset; their derivation through calculations involving certain major ions and the hydrological characteristics they share with some of the exogenous parameters included in the research, such as precipitation, rather than being solely determined by the probability of correlating with any of the water quality parameters.

For example, the presence of ALKT in water bodies is primarily attributed to the dissolution of minerals from the surrounding rocks and land when precipitation occurs. During runoff events, it gathers chemicals like Calcium Carbonate (CaCO3) and introduces them into the water body (Omernik et al., 2018). On the other hand, HARD signifies the concentration of dissolved calcium and magnesium in water sources. The high data availability in past years allows for better imputation of missing values using state-of-the-art imputation algorithms. The relationship between water quality parameters is different in different water systems (Saalidong et al., 2022); therefore, this study did not make any assumptions and did not investigate the correlation between these parameters.

Based on the domain knowledge gathered, TDS also has organic ions, namely Carbonates (CO_3^{2-}) and Bicarbonates (HCO_3^-), that were not recorded by the monitoring programs.

However, these values were calculated using ALKT and pH and served as supporting parameters to model the predictions of major ions. Carbonate and bicarbonate are important ions in water chemistry. The concentration of these ions can provide information about the water's pH, ALKT, and HARD. Here are the equations for calculating carbonate and bicarbonate concentrations in water:

1. Carbonate (CO_3^{2-}) Calculation:

$$CO_3^{2-} = \frac{Total\ Alkalinity}{10^{pH-pKa_1} + 10^{pH-pKa_2}} \qquad \textit{Equation 2}$$

where pK_{a1} and pK_{a2} are the dissociation constants for Carbonic Acid (H_2CO_3). The values of pK_{a1} and pK_{a2} are 6.35 and 10.33, respectively, at 25°C (Averill & Eldredge, 2011).

2. Bicarbonate (HCO_3^-)

$$HCO_3^- = \frac{Total\ Alkalinity\ - CO_3^{2-}}{1 + 10^{pKa_1-pH}} \qquad \textit{Equation 3}$$

It is important to note that these equations assume that only carbonate, bicarbonate, and hydroxide ions contribute to the alkalinity of the water. Other ions, such as phosphate, silicate, borate, and organic acids, may also contribute to alkalinity and affect the accuracy of these calculations. Fortunately, there are no significant measures for these parameters, such as phosphate and silicate, in the database; hence, it allowed for carbonates bicarbonates to be calculated and incorporated into the water quality modelling during

the feature engineering process. Feature engineering is the process of selecting, transforming, or creating new features (input variables) from the raw data to improve the performance of a machine learning model.

Furthermore, to mitigate data scarcity, this study integrates hydrometric and meteorological data with water quality data to explore the possibility of improving the predictive ability of constructed models. These include Atmospheric Mean Temperature ($temp_a$), Total Precipitation ($precip_{ttl}$), Stage, Discharge, Relative Humidity ($humid_{rel}$), Wind Direction and Wind Gust.

2.4 Knowledge Gaps and Limitations

Several insights were gained from the presented literature review. First, the literature review reveals a significant gap in addressing sparsity in water quality datasets. Many reviewed papers should have focused on handling missing values or effectively dealing with sparse data.

Second, the oversight in feature selection is apparent. The researchers chose water quality parameters primarily based on their availability rather than considering their interdependencies. When selecting parameters for data-driven modelling approaches, accounting for the interactions among parameters within water bodies is crucial. For instance, careful consideration is necessary when selecting constituents profoundly influenced by biological processes in water. This aspect necessitates further investigation into the underlying biological processes.

Third, the analysis in some papers should have considered the entirety of the watershed system, which can lead to overlooking crucial parameters that may play a significant role in determining water quality.

Fourth, there was no inclusion of exogenous parameters such as hydrological and weather parameters, which are known to influence water quality in natural water bodies.

Additionally, it uncovered valuable insights into the challenges of applying process-based models to natural water resources. One major limitation is the complexity of modelling water quality in watersheds due to the diverse factors affecting it, including anthropogenic activities fueled by population growth, precipitation, vegetation, depositions, and climate change[1]. The water science team at Environment and Climate Change Canada (ECCC) utilizes sophisticated mathematical models and supercomputers to comprehend the impacts of various factors on water quality. Furthermore, several challenges remain, particularly concerning the modelling of water movement across different landscapes, such as farms, forests, grasslands, lakes, and river basins. Also, understanding the dynamics of water flow in these diverse environments is crucial for comprehending water quality variations (ECCC, 2022).

These complexities emphasize the need for data-driven approaches like deep learning models to address (1) the challenges posed by sparsity in water quality data and (2) the expense of using process-based models. Such models can approximate multivariate time series observations by leveraging routinely collected data from monitoring systems. The advantage of such approximation is that such models need only values of concentrations of water constituents and maybe other endogenous and exogenous features. On the contrary, process-based models need the values of observed concentrations along with the model parameters that describe the speed of processes included in the model. It is not easy to obtain accurate values of such model parameters.

---

[1] Processes to consider in a model: https://www.deq.nc.gov/water-quality/planning/tmdl/modeling/modeling-101-fon-stakeholder-may09/download

2.5 Scope of the Study and Research Questions

The primary objective of this research is to develop models for predicting water quality in a small watershed in a densely populated area. The focus will be exploring various deep learning architectures and approaches to achieve accurate predictions.

Furthermore, the research aims to create a comprehensive framework that can be utilized for modelling water quality in complex natural resources, even when faced with challenges such as high sparsity and intermittency in environmental datasets.

By employing deep learning techniques and creating a robust framework, this study endeavours to contribute to water quality modelling, providing valuable insights for water resource management and decision-making in similar settings.

To address the limitations in section 2.4, various techniques for handling missing values, such as imputation methods specifically designed for time series data, are employed. Feature engineering is carefully considered to capture temporal patterns and pertinent relationships in the water quality parameters.

Furthermore, a holistic approach that encompasses the entire watershed system is adopted that includes relevant parameters and their interactions to provide a more detailed description of water quality dynamics and to help create more robust and accurate predictive models. By addressing these gaps, the study will contribute to more effective water resource management and decision-making processes.

Therefore, the goal is to answer the following research questions:

- RQ 1: Which data-driven model provides reliable and efficient predictions of major ions in natural resources?
- RQ 2: How do imputed values impact the data-driven water quality prediction framework?

- RQ 3: How do exogenous parameters impact the data-driven water quality prediction framework?

3 Methodology

3.1 Knowledge Discovery in Databases and the Proposed Framework

Commencing from the inception of a research concept and culminating in the completion of the final report, the research process encompasses a series of pivotal stages (Bordens & Abbott, 2022).

Primarily, the impetus to explore the modelling of water quality in natural resources emerged as a response to the knowledge gaps and limitations elucidated in section 2.4. A conceptual research process to develop the water quality modelling framework (Figure 7) was first designed. Within each step of the research process, sub-routines were developed to prepare, test, and evaluate the predictive models investigated in the research (Figure 8).

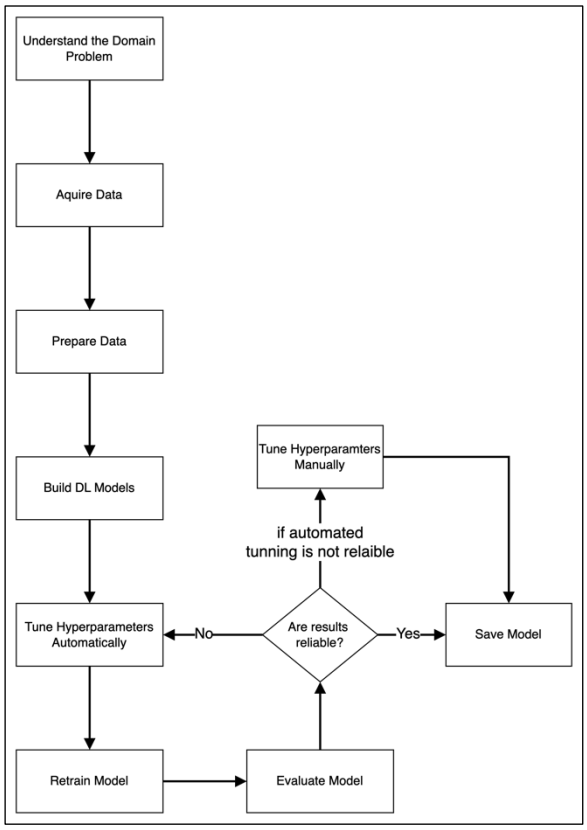


Figure 7 A conceptual research process

After a comprehensive understanding of the problem domain and its underlying concepts, the research began with extracting data from open sources using APIs implemented in Python (step 1). Step 2 involved cleansing the data by removing erroneous data (recorded as -99 in the datasets). Step 3 involved segmenting the datasets into seven predefined subwatersheds, which guarantees the preservation of parameter interdependencies within each subwatershed.

In the Don River Watershed, there are three major watersheds: the West Don, the East Don and the Lower Don and within these three are seven subwatersheds (Figure 14).

From steps 4 to 6, the datasets were aggregated and resampled to monthly frequencies. This process started from the upper subwatersheds to the lower subwatershed within a major watershed. For instance, the Upper West Don was aggregated with the Lower West Don data and then resampled to monthly means to (1) remove duplicates and (2) augment the missing values with available data in any datasets. Afterwards, the water quality dataset was merged instantaneously (date for date) with the hydrometric and meteorological datasets to form a single dataset for each major watershed. To augment the missing values in the Lower Don, the Taylor/Massey Creek dataset was combined with the Lower Don's dataset.

Steps 7 to 9 involved imputing missing values using different imputation techniques, section 4.1.2.

During steps 10 to 12, additional features were engineered, including Water Quality Indices (WQI), indicators for failed water quality guidelines, missing value indicators, and major watershed indicators. The Water Quality Index (WQI) is a numerical representation to evaluate and communicate the comprehensive state of water quality within a specific environment. It is a composite metric that considers various attributes such as physical, chemical, and biological aspects, aggregating them into a single value that depicts the overall condition of a water body.

The indicators for failed water quality guidelines are binary values (0s and 1s) used to signify whether a particular water quality parameter has failed (1) to meet established guidelines or not (0). These guidelines adhere to the directives outlined by the Canadian Council of Ministers of the Environment (CCME).

Similarly, the missing value indicators denote whether a value has been imputed (1) or not (0). These indicators contribute to understanding the data imputation process.

The major watershed indicators employ a one-hot-shot encoding to differentiate among the three watersheds. In this encoding, a value of "1" signifies the source location of the samples, as the major watershed datasets were merged before the modelling phase.

It is worth noting that this research proposes a modelling framework that explores two distinct approaches to ascertain the preferable predictive method. The first approach assumes that significant variations in major ion concentrations from the upstream and downstream sections of a river are not reflected in the available data sets. Consequently, the entire watershed system can be treated as a singular entity for predicting major ion concentrations. Conversely, the second approach posits that water quality from the upstream locations influences concentrations of the investigated water constituents in the downstream river cross-section, allowing upstream observations to forecast downstream water quality.

In step 12, the dataset is split into training and test sets with a 75% and 25% ratio, respectively. In order to preserve the time sequence of ordered months, the data was split in a time-series fashion where the most recent months were in the test set. Batches of data from the past month () along with a target value (concentration of the major ion in perspective) in the future () using an in-built Keras utility, namely `timeseries_dataset_from_array`(), were created. When working with time series data, it is important to use validation and test data that are more recent than the training data. Because future water quality concentrations are being predicted given the past, the validation/test split must reflect this notion, thus preserving the temporal sequence inherent in time series data.

Step 13 involved the scaling of the data. In the case of this research, outliers or extreme values were important to investigate; hence, the standard scaler (Equation 15) was selected for scaling. Scaling helps to bring all the data values to equal ranges before modelling.

The last step, step 14, involved the modelling of the data. This step has sub-steps such as the hyperparameter tuning, model evaluation and model testing. Hyperparameter tuning involves finding the best hyperparameters that aid in finding suitable model results. Hyperparameter tuning can be done automatically using available Python libraries such as Keras Tuner or Optuna. The Optuna hyperparameter optimization library was used. OT is a specialized library designed to facilitate the selection of the most appropriate hyperparameter configuration for most ML and DL models. The choice of OT was driven by its alignment with TensorFlow, making it the preferred choice due to its compatibility, speed, and easy integration with most deep learning frameworks. The hyperparameters tuned were the number of layers and units in a NN, the percentage of dropouts, the learning rate, the number of epochs and the activation function to choose from. A manual hyperparameter tuning is required if results from the automation are not satisfying.

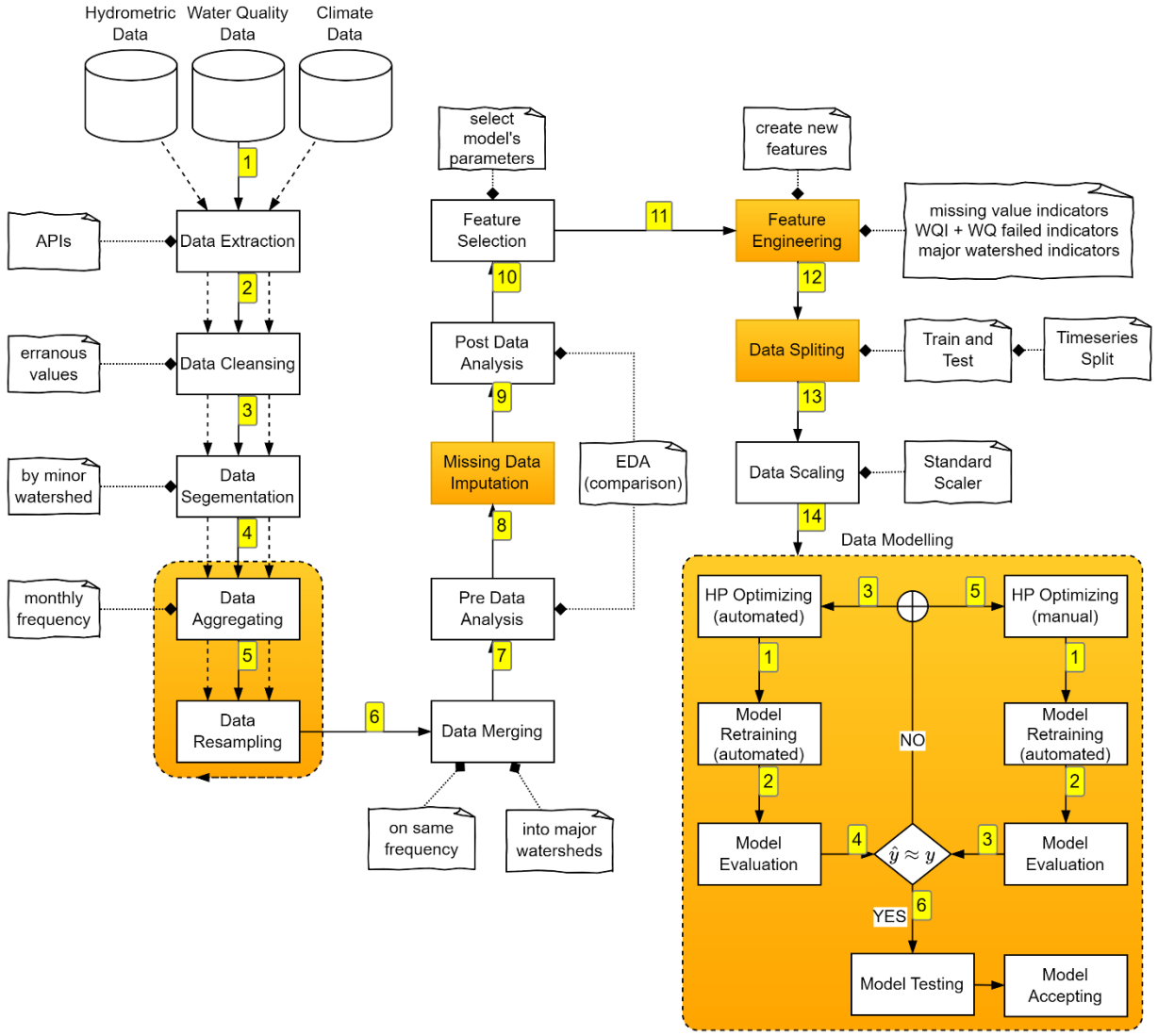


Figure 8 Data-driven water quality prediction framework.

Activation functions included in this study were Rectified Linear Unit (ReLU) (Nair & Hinton, 2010), Exponential Linear Unit (ELU) (Clevert et al., 2015), Scaled Exponential Linear Unit (SELU) (Klambauer et al., 2017), Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel, 2016) and Hyperbolic Tangent (tanh) functions.

The ReLU function is defined as:

$$ReLU(x) = max(0, x)$$ *Equation 4*

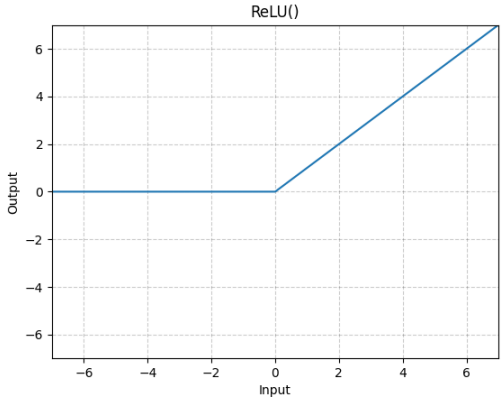
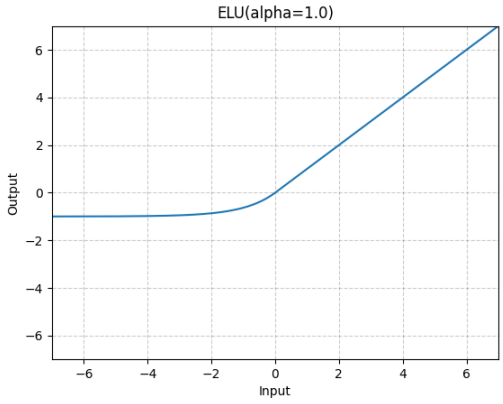


Figure 9 Rectified Linear Unit function[2]

The ELU function is defined as:

$$ELU(x) = \begin{cases} x, & if\ x > 0 \\ \alpha \cdot (e^x - 1), & if\ x \leq 0 \end{cases}$$ *Equation 5*



[2] Source: ReLU — PyTorch 2.0 documentation

Figure 10 Exponential Linear Unit function[3]

The SELU function is defined as:

$$SELU(x) = \lambda \begin{cases} x, & if\ x > 0 \\ \alpha \cdot (e^x - 1), & if\ x \leq 0 \end{cases} \qquad \textit{Equation 6}$$

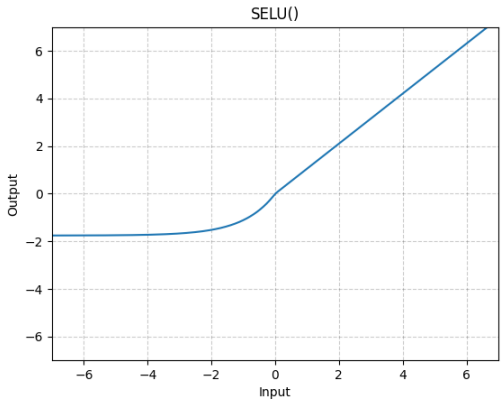


Figure 11 Scaled Exponential Linear Unit function[4]

The GELU function is defined as:

$$GELU(x) = x \cdot \phi(x) \qquad \textit{Equation 7}$$

[3]Source: ELU — PyTorch 2.0 documentation

[4]Source: SELU — PyTorch 2.0 documentation

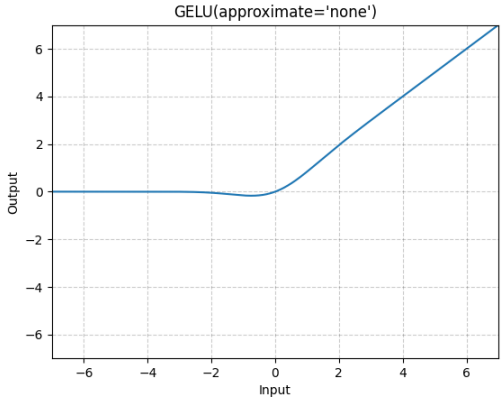


Figure 12 Gaussian Error Linear Unit function[5]

where $\phi(x)$ is the cumulative distribution function for gaussian distribution.

The Tanh function is defined as:

$$tanh(x) = \frac{e^x - e^x}{e^x + e^x} \qquad \textit{Equation 8}$$

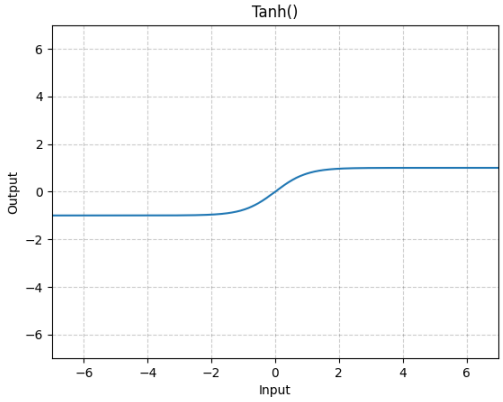


Figure 13 Hyperbolic Tangent function[6]

During the model's training, 25% of the train set was held out for validation. Again, the data needed to be shuffled to preserve the time-sequence order in the data.

---

[5]Source: GELU — PyTorch 2.0 documentation

[6]Source: Tanh — PyTorch 2.0 documentation

The loss functions used were the Mean Squared Error (MSE); this is great for ensuring that our trained model can detect outliers or extreme values since the MSE puts a larger weight on these errors due to the squaring part of the function, and the Mean Absolute Error (MAE); this takes the absolute value which weights all the errors on the same linear scale. Unlike the MSE, this loss function does not weigh too much on outliers or extreme values and provides a generic and even measure of how well our model performs.

After training, the model is re-trained on the full training set (train and validation set put together) before being evaluated on the test set. Predictions are made on the test set as well to finally see how well the model will generalize on other similar datasets.

3.2 Software Selection

Data science has gained immense popularity due to the rapid expansion of open-source software for knowledge discovery and DL, accompanied by numerous online tutorials. Python, R, MATLAB, SAS, SPSS, and STATA are the most popular ones (Brittain et al., 2018; Colaco et al., 2020).

(Ozgur et al., 2021) studied several data science tools and reported that Python was the most easily learned, read, and used programming tool. Python version 3.10.6 and Google's Phyton Integrated Development Environment, Colab, were chosen for the water quality modelling in this study. Libraries including Pandas, NumPy, Scikit-learn, Keras-Tuner, Optuna, TensorFlow/Keras, Lets-Plot, and Matplotlib were used in the context of this research.

TensorFlow was selected for creating, modelling, and testing the predictive models. It has a large community support with well-documented frameworks and a large repository of trained models and tutorials. TensorFlow also beats most DL packages in deploying the trained model to production. In this study, TensorFlow version 2.13.0 was used.

Additionally, ArcGIS was selected for geospatial analysis and used to map out the watershed, its subwatersheds, river courses and locations of the monitoring sites[7].

3.3 Data Acquisition

The Don River Watershed (Figure 14) s a significant natural object in Toronto, Ontario, Canada. It encompasses a vast area spanning approximately 361 square kilometres (about half the area of Austin, Texas) and serves as an essential ecological and hydrological system within Toronto. The Don River, the watershed's central watercourse, originates from the Oak Ridges Moraine, a geological formation north of Toronto. It flows southward, traversing various urban and natural landscapes before emptying into Lake Ontario. Despite these facts, the water discharge of the river could be a lot higher and is subjected to high rapid increases under intensive precipitations. A diverse range of land uses, including residential, industrial, commercial, and recreational areas, characterizes the watershed. The Don River Watershed has experienced significant modifications and urbanization as Toronto has developed. According to the 2018 report (the most recent report as of the time of this thesis), 85% of the land use in the Don River watershed is completely urban, 14% is natural cover with less than 2% rural land cover (TRCA, 2018). Figure 15 depicts the main subwatersheds of the Don River with all the monitoring stations used in this research.

1. [7] Area map of the Don River watershed: https://arcg.is/01yDCP

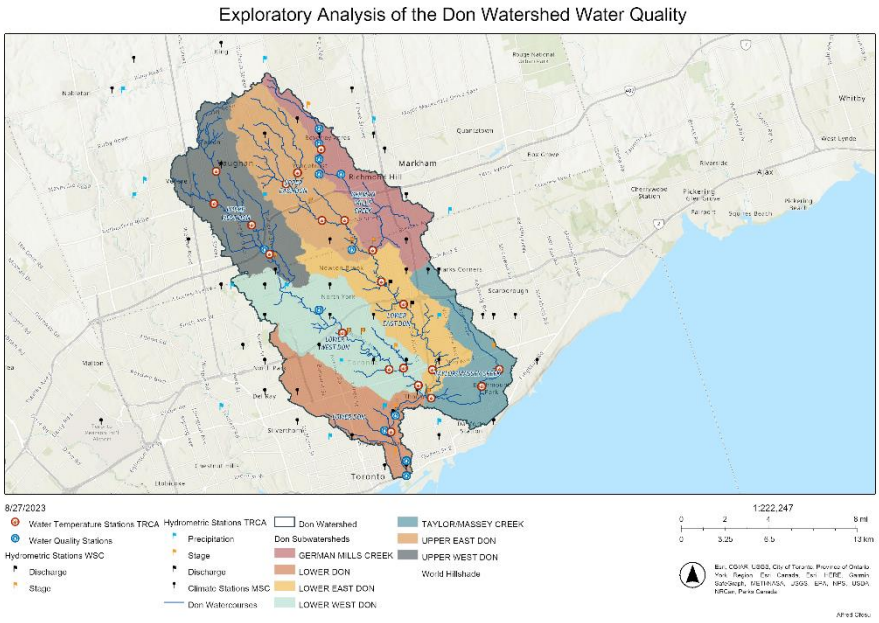


Figure 14 Don River Watershed with Monitoring Stations[8]

The Toronto and Regional Conservation Authority (TRCA) monthly monitors water quality in the Don River watershed. This data is then reported to the Ministry of the Environment, Conservation and Parks (MECP) under the Provincial Water Quality Monitoring Network (PWQMN) program and to the public. The water quality parameters were collected from the PWQMN's and TRCA's databases from 1964 to 2021. In total, there are 2119 number of observations and over 80 water quality parameters. Hydrometric data, including stream levels, precipitation, and discharges, were obtained from TRCA's database. Weather data, such as atmospheric temperature, precipitation, and wind speed, were also procured from the Meteorological Service Canada (MSC) for the same period as the water quality data. Likewise, additional hydrometric data was obtained from the Water Survey of Canada (WSC). The datasets from the MSC and WSC

[8] source: Ofosu, Alfred

are recorded using automated systems that capture fluctuations every 15 minutes and send the reading to their databases.

Some of the commonly monitored water quality parameters in this watershed include Ammonia (NH3), Bacteria (including fecal coliform and E. coli), Chloride (Cl-), Conductivity (a measure of the ability of water to conduct an electrical current), Dissolved oxygen (DO), Metals (including lead, mercury, and arsenic), Nitrate (NO3-), Nutrients (nitrogen and phosphorus compounds), pH (acidity or basicity), Sulphate (SO42-), Temperature, Total coliform bacteria, Total nitrogen (TN), Total organic carbon (TOC), Total phosphorus (TP), Total suspended solids (TSS), Turbidity (a measure of water clarity), and many more.

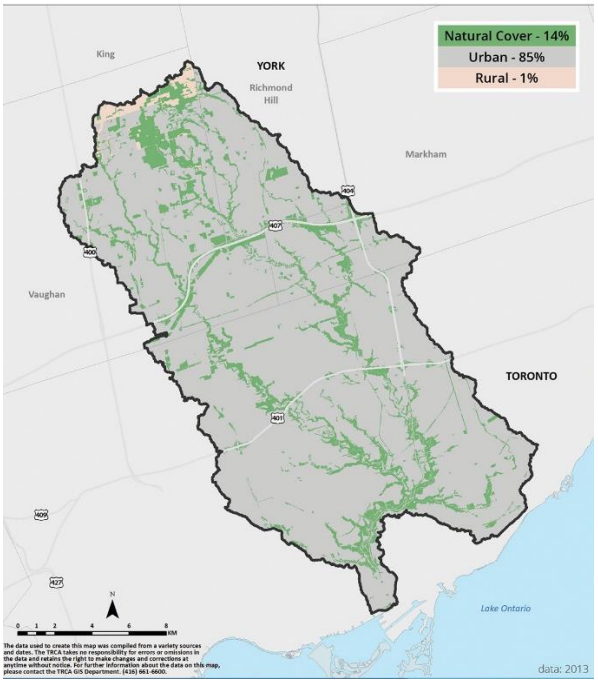


Figure 15 Don River Watershed with Landcover - source: 2018 Report Card, TRCA

The water quality parameters selected for modelling were calcium (Ca2+), chloride (Cl-), potassium (K+), magnesium (Mg2+), sodium (Na+), sulphate (SO42-), total dissolved solids (TDS), Alkalinity (ALKT), Electric Conductivity (CNDT), hardness (HARD), pH and water temperature ($temp_w$). The meteorological parameters selected for modelling

were precipitation (precip), mean atmospheric temperature ($temp_a$), wind gust, wind direction and mean relative humidity (mean_rel_humidity). The wind direction was in units of degrees, and angles may not be good model inputs, so in combination with the wind gust, they were converted into a wind vector for the model to interpret. The hydrometric parameters selected for modelling were stage (river level in m) and discharge (river flow rate/s).

A primary contributor to the inaccuracies encountered in modelling rainfall and temperature lies in the error stemming from the spatial heterogeneity of these hydrological parameters (TRCA & AECOM, 2017). A network of monitoring stations was strategically chosen around and within each subwatershed to effectively encompass the spatial and temporal dispersion of rainfall and temperature - encompassing both water and atmospheric conditions - across the Don River watershed. This approach aims to model the mean values for each parameter, thus preventing the model from presuming that the measured hydrological parameter at a particular location adequately represents the entirety of rainfall and temperature occurring over the entire study area.

3.4 Monitoring Stations

The geographic locations of the monitoring stations used in this study are depicted in Figure 14, showcasing their distribution within and around the Don River watershed. The amalgamated dataset encompasses more than five decades of data records. A comprehensive summary of all 102 monitoring stations is presented in Table 2 in the appendix. The study encompasses seven subwatersheds: Upper Don West, Upper Don East, German Mills Creek, Lower Don West, Lower Don East, Taylor/Massey Creek, and Lower Don. Water quality parameter values were collected by the PWQMN from 1954 to approximately 1999 for all subwatersheds. Observations for the Lower Don subwatershed recommenced in the year 2000, whereas data collection began in 2015 for

the other subwatersheds. The subsequent sections illustrate the concentrations of major ions in each subwatershed. Significantly, a consistent pattern becomes evident when examining monthly averages, with the fluctuations in upstream values closely mirroring those observed downstream. This validates the assumption underpinning the first modelling approach.

3.4.1 Upper West Don Subwatershed

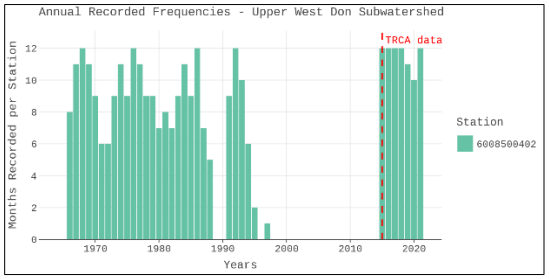


Figure 16 Recorded frequencies of water quality parameters in the Upper West Don subwatershed.

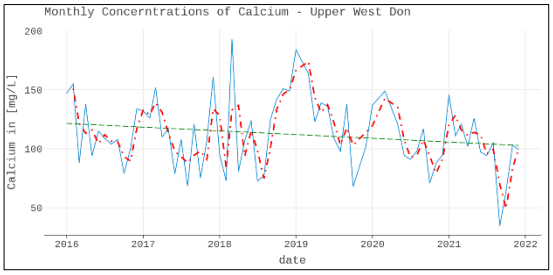


Figure 17 Calcium concentrations in the Upper West Don subwatershed

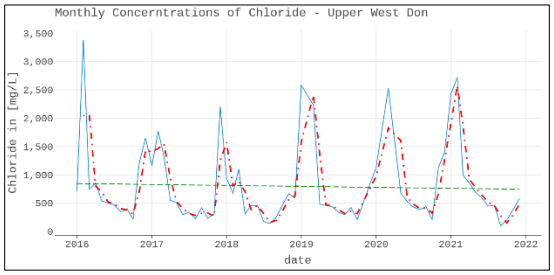


Figure 18 Chloride concentrations in the Upper West Don subwatershed

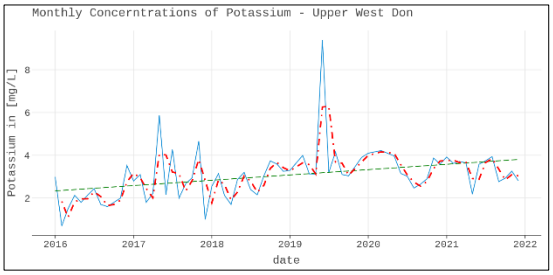


Figure 19 Potassium concentrations in the Upper West Don subwatershed

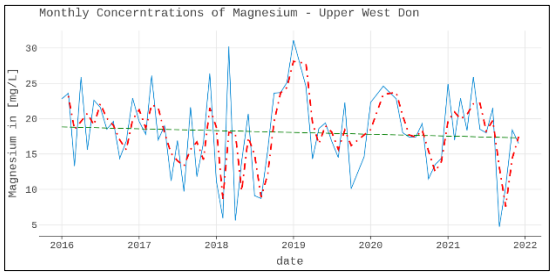


Figure 20 Magnesium concentrations in the Upper West Don subwatershed

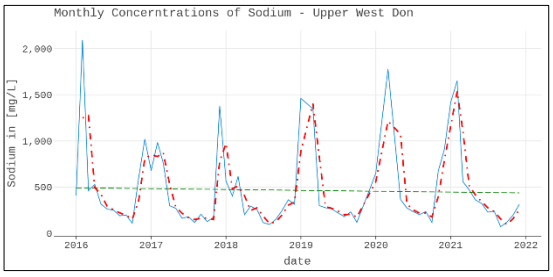


Figure 21 Sodium concentrations in the Upper West Don subwatershed

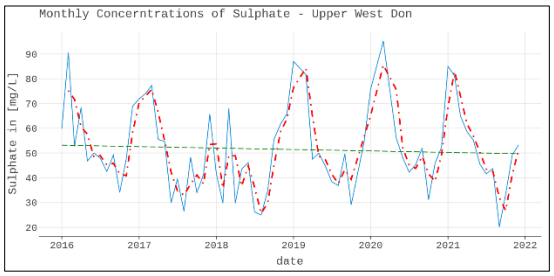


Figure 22 Sulphate concentrations in the Upper West Don subwatershed

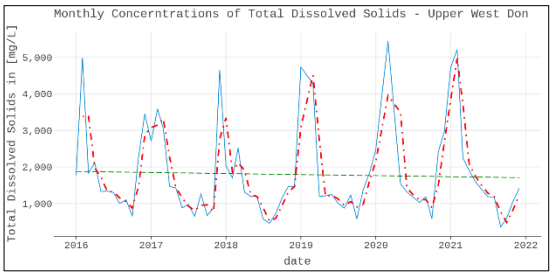


Figure 23 TDS concentrations in the Upper West Don subwatershed

The Upper West Don subwatershed is equipped with a singular water quality monitoring station, namely 6008500402. This station's operation spans from 1964 to 2021; however, it exhibits intermittent gaps in observations during the early 1990s and then from approximately 2000 to 2015. Notably, the most robust collection of constituent data transpired between 2015 and 2021, coinciding with the Toronto and Regional Conservation Authority (TRCA) initiating water quality monitoring in this subwatershed. An analysis of the major ion concentrations reveals consistent fluctuations over these years. Particularly, there is a minor upward trend in potassium levels, with an approximate increase of 2 mg/L from 2015 to 2021.

3.4.2 Lower West Don Subwatershed

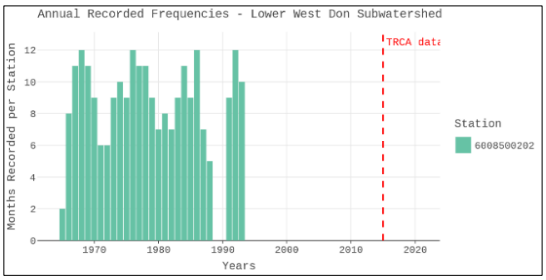


Figure 24 Recorded frequencies of water quality parameters in the Lower West Don subwatershed.

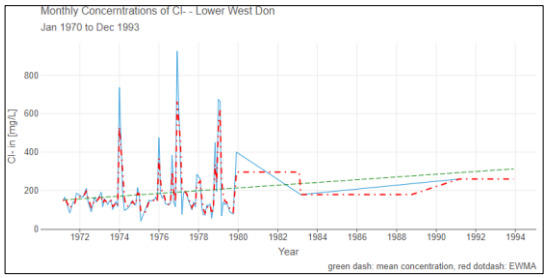


Figure 25 Chloride concentrations in the Lower West Don subwatershed

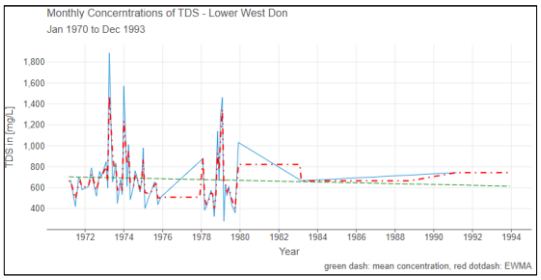


Figure 26 TDS concentrations in the Lower West Don subwatershed

This watershed had only two major ions, Chloride and TDS, which were then aggregated with the Upper West Don subwatershed data to augment the other non-available concentrations.

3.4.3 Upper East Don Subwatershed

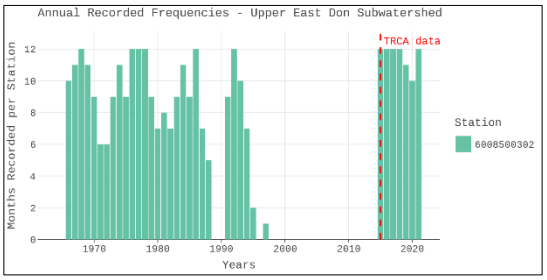


Figure 27Recorded frequencies of water quality parameters in the Upper East Don subwatershed.

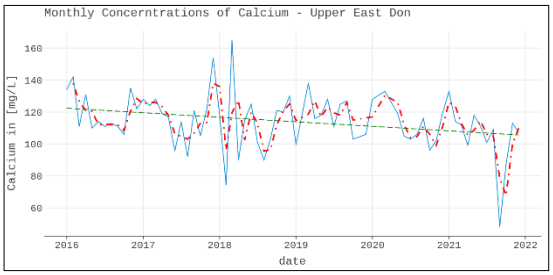


Figure 28 Monthly Calcium concentrations in the Upper East Don subwatershed

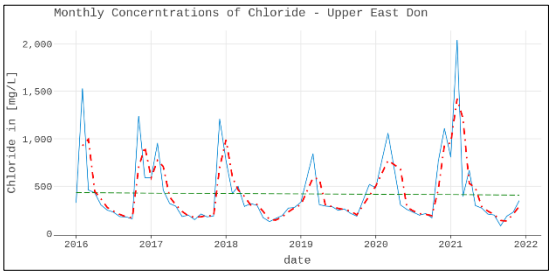


Figure 29 Monthly Chloride concentrations in the Upper East Don subwatershed

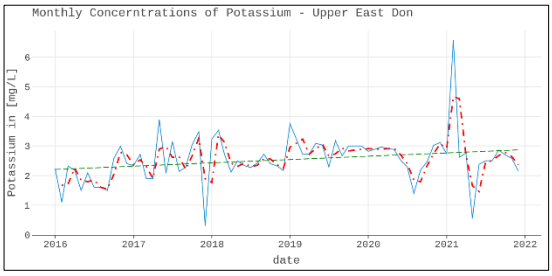


Figure 30 Monthly Potassium concentrations in the Upper East Don subwatershed

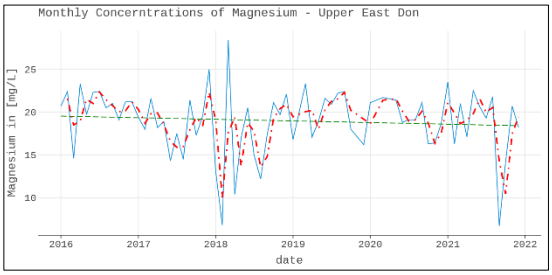


Figure 31 Monthly Magnesium concentrations in the Upper East Don subwatershed

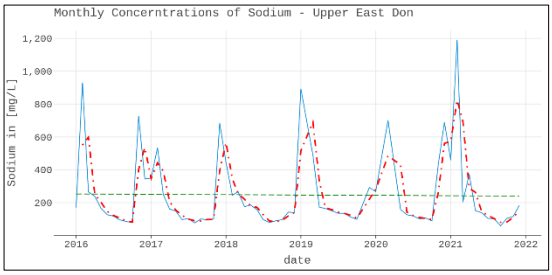


Figure 32 Monthly Sodium concentrations in the Upper East Don subwatershed

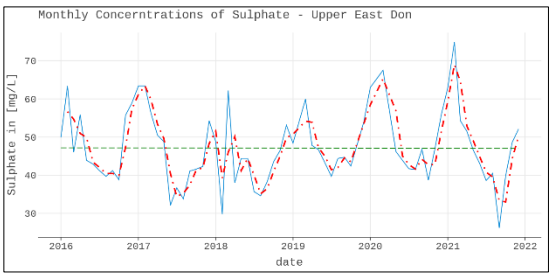


Figure 33 Monthly Sulphate concentrations in the Upper East Don subwatershed

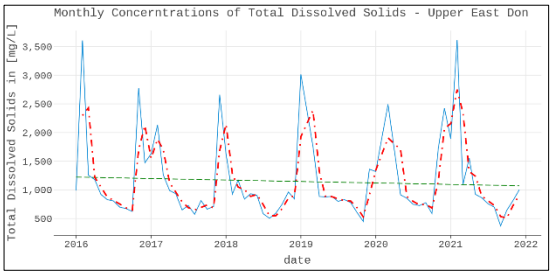


Figure 34 Monthly TDS concentrations in the Upper East Don subwatershed

Within this subwatershed is another water quality monitoring station, denoted as 6008500302. Like station 6008500402, this station also exhibits gaps in observations. Additionally, potassium concentrations display a slight upward trend from 2015 to 2021. However, after analyzing its monthly means over time, more than this increase is needed to warrant the application of a stationary transformation.

3.4.4 Lower East Don Subwatershed

While this subwatershed lacks dedicated water quality monitoring stations, it possesses hydrometric and meteorological stations. To address this gap and enhance the overall water quality modelling in the East Don major watershed, the data from these hydrometric and meteorological stations were amalgamated with the data from the Upper East Don subwatershed. This approach aims to provide a more comprehensive understanding of the water quality in the East Don major watershed.

3.4.5 German Mills Creek Subwatershed

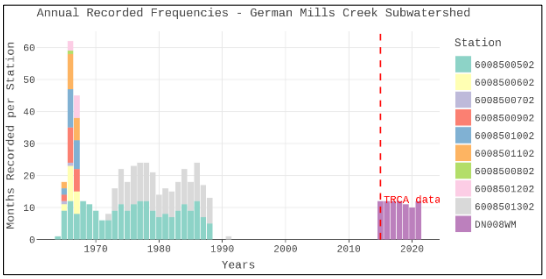


Figure 35 Recorded frequencies of water quality parameters in the German Mills Creek subwatershed.

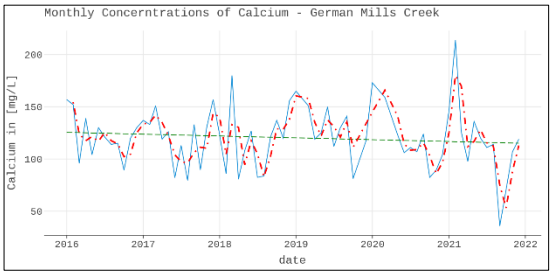


Figure 36 Monthly Calcium concentrations in the German Mills Creek subwatershed

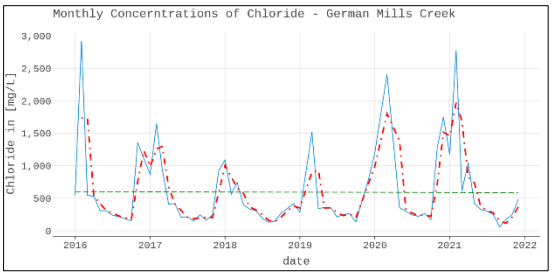


Figure 37 Monthly Chloride concentrations in the German Mills Creek subwatershed

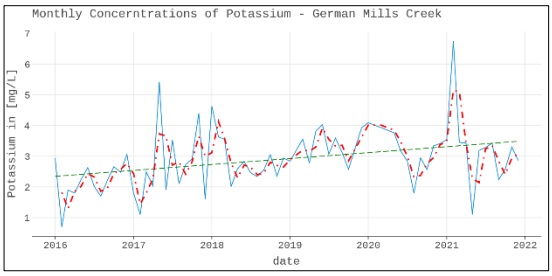


Figure 38 Monthly Potassium concentrations in the German Mills Creek subwatershed

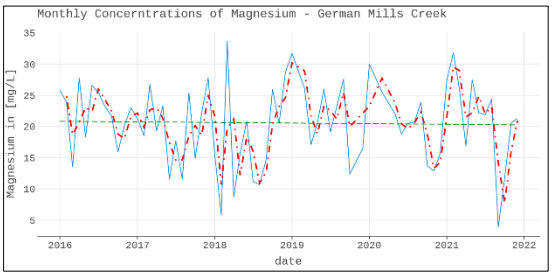


Figure 39 Monthly Magnesium concentrations in the German Mills Creek subwatershed

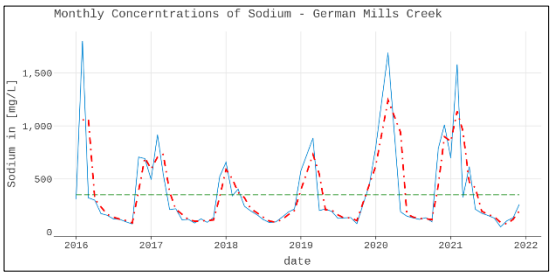


Figure 40 Monthly Sodium concentrations in the German Mills Creek subwatershed

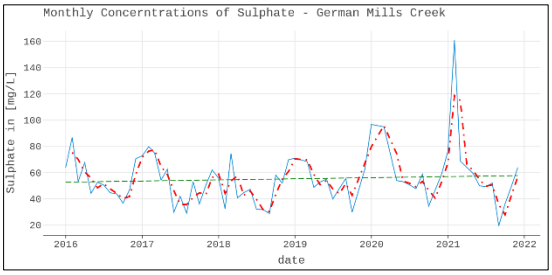


Figure 41 Monthly Sulphate concentrations in the German Mills Creek subwatershed

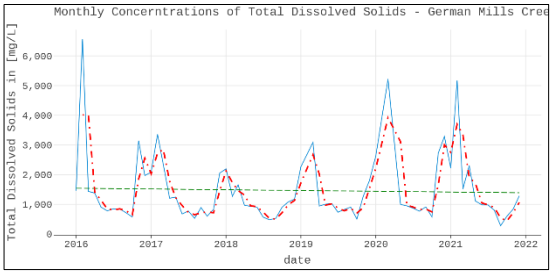


Figure 42 Monthly TDS concentrations in the German Mills Creek subwatershed

The German Mill Creek subwatershed is situated northeast of the Don River watershed and has headwaters from the Oak Moraine Ridges. It has the highest number of water quality stations, namely 6008500502, 6008500602, 6008500702, 6008500802, 6008500902, 6008501002, 6008501102, 6008501202, 6008501302, and DN008WM. Water quality was monitored at nine stations from 1964 to about 1988. Observations were discontinued until

2015 when monitoring began at the DN008WM station. Potassium shows a slight upward trend in the watershed as well.

3.4.6 Taylor/Massey Creek Subwatershed

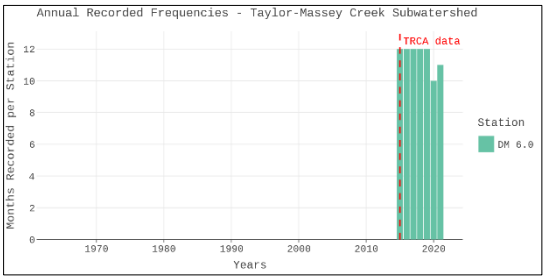


Figure 43 Recorded frequencies of water quality parameters in the Taylor/Massey Creek subwatershed.

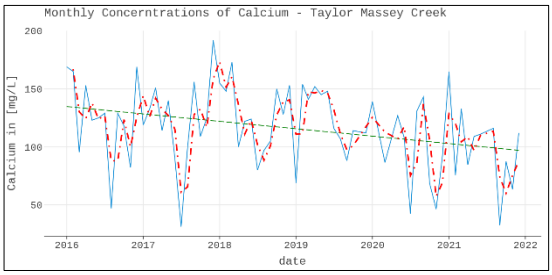


Figure 44 Monthly Calcium concentrations in the Taylor/Massey Creek subwatershed

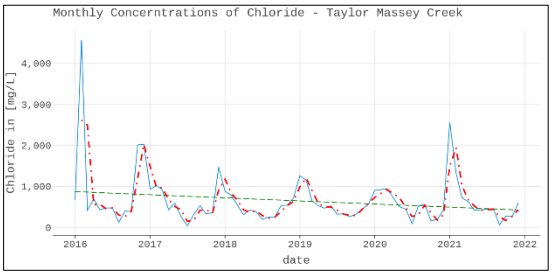


Figure 45 Monthly Chloride concentrations in the Taylor/Massey Creek subwatershed

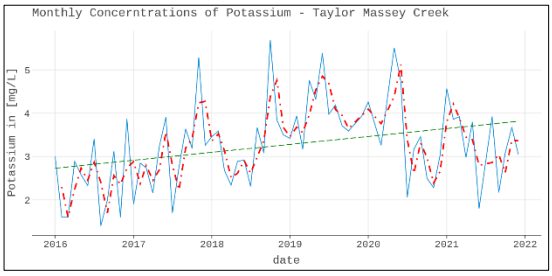


Figure 46 Monthly Potassium concentrations in the Taylor/Massey Creek subwatershed

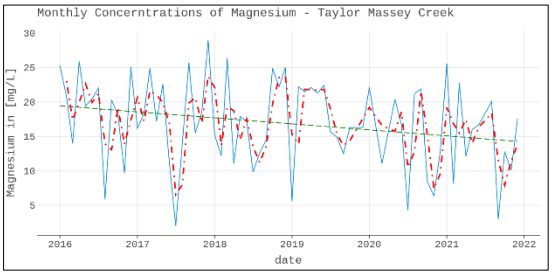


Figure 47 Monthly Magnesium concentrations in the Taylor/Massey Creek subwatershed

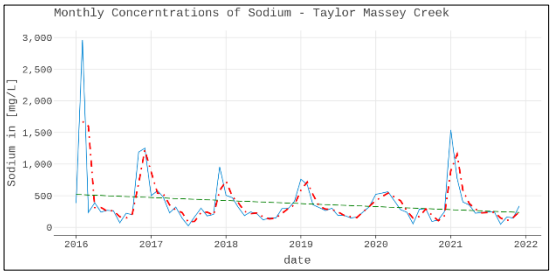


Figure 48 Monthly Sodium concentrations in the Taylor/Massey Creek subwatershed

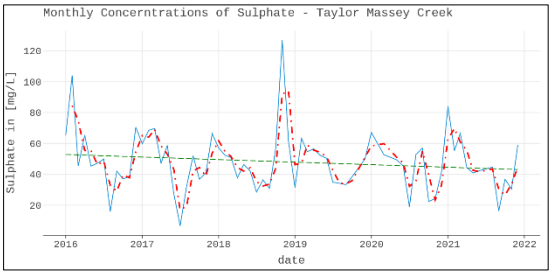


Figure 49 Monthly Sulphate concentrations in the Taylor/Massey Creek subwatershed

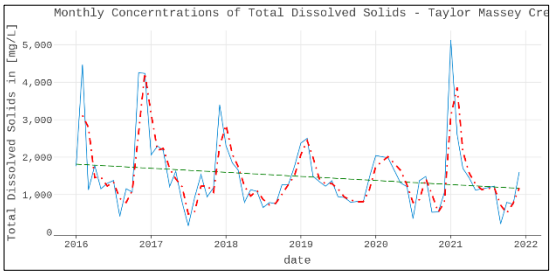


Figure 50 Monthly TDS concentrations in the Taylor/Massey Creek subwatershed

The Talyor/Massey Creek subwatershed is located southeast of the Don River watershed and has only one water quality monitoring station, DM 6.0, which began service in 2015. In this subwatershed, it is noticeable that potassium has an upward trend, whereas magnesium has a downward trend. This is significantly small, and after analyzing its monthly means over time, this won't require a stationary transformation.

3.4.7 Lower Don Subwatershed

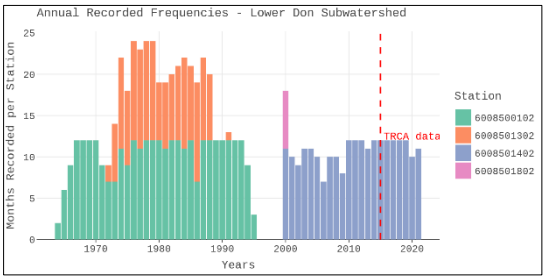


Figure 51 Recorded frequencies of water quality parameters in the Lower Don subwatershed.

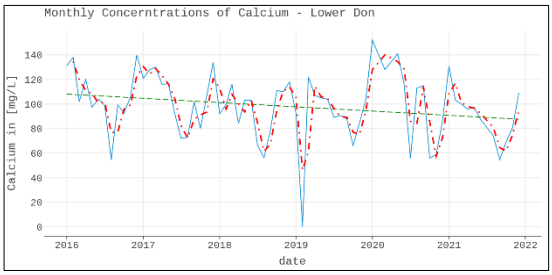


Figure 52 Monthly Calcium concentrations in the Lower Don subwatershed

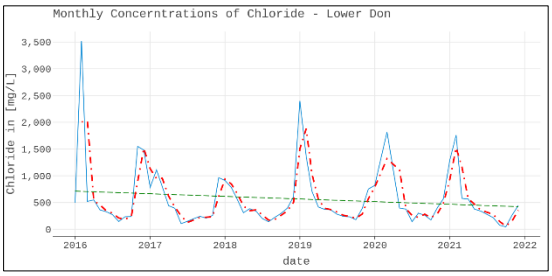


Figure 53 Monthly Chloride concentrations in the Lower Don subwatershed

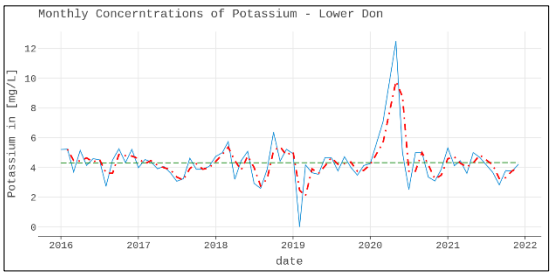


Figure 54 Monthly Potassium concentrations in the Lower Don subwatershed

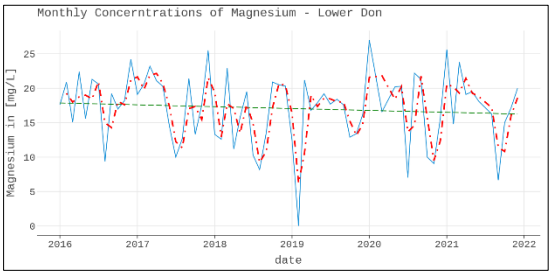


Figure 55 Monthly Magnesium concentrations in the Lower Don subwatershed

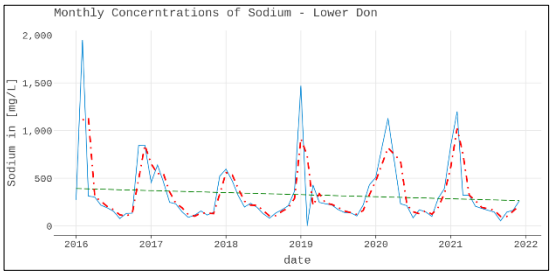


Figure 56 Monthly Sodium concentrations in the Lower Don subwatershed

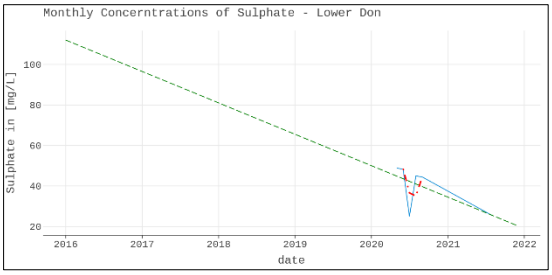


Figure 57 Monthly Sulphate concentrations in the Lower Don subwatershed

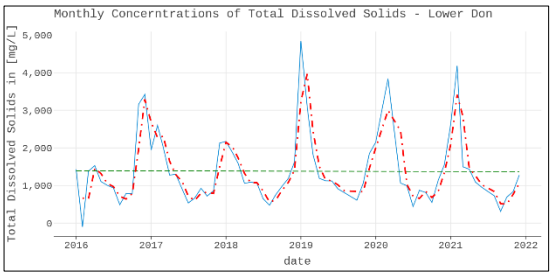


Figure 58 Monthly TDS concentrations in the Lower Don subwatershed

Within the Lower Don subwatershed, four monitoring stations are identified as 6008500102, 6008501302, 6008501402, and 6008501802. This subwatershed displays fewer gaps in observations compared to the monitoring stations in other subwatersheds. Notably, chloride concentrations have demonstrated a decreasing trend over the past seven years. Conversely, sulphate has been detected in only four months throughout the seven years. This limited occurrence renders it inadequate for effective water quality modelling within the Lower Don subwatershed.

For the missing values of Sulphate concentration in the Lower Don, the Talyor/Massey Creek dataset was aggregated with the Lower Don dataset to augment the missing values. If there were any additional missing values, they were imputed using imputation techniques. Regarding this approach, Taylor/Massey Creek was combined with the Lower Don to form the Lower Don major watershed.

4 Results and Discussion

4.1 Data Transformation

4.1.1 Time Series Data

Time series data form a sequence of data points indexed or ordered based on time. Normally, such data represent a collection of observations or measurements recorded at different points in time. Time series data can be used to analyze and predict evolving trends, patterns, and behaviour. This research collected monthly observations (e.g., Figure 59) of water quality parameters from 6 of 7 subwatersheds. Each subwatershed provided a multivariate time series that allowed the inclusion of all or some of them into a set of the model's features jointly. Additionally, as the datasets were combined, it provided the models with multiple time series, exposing them to various observations.

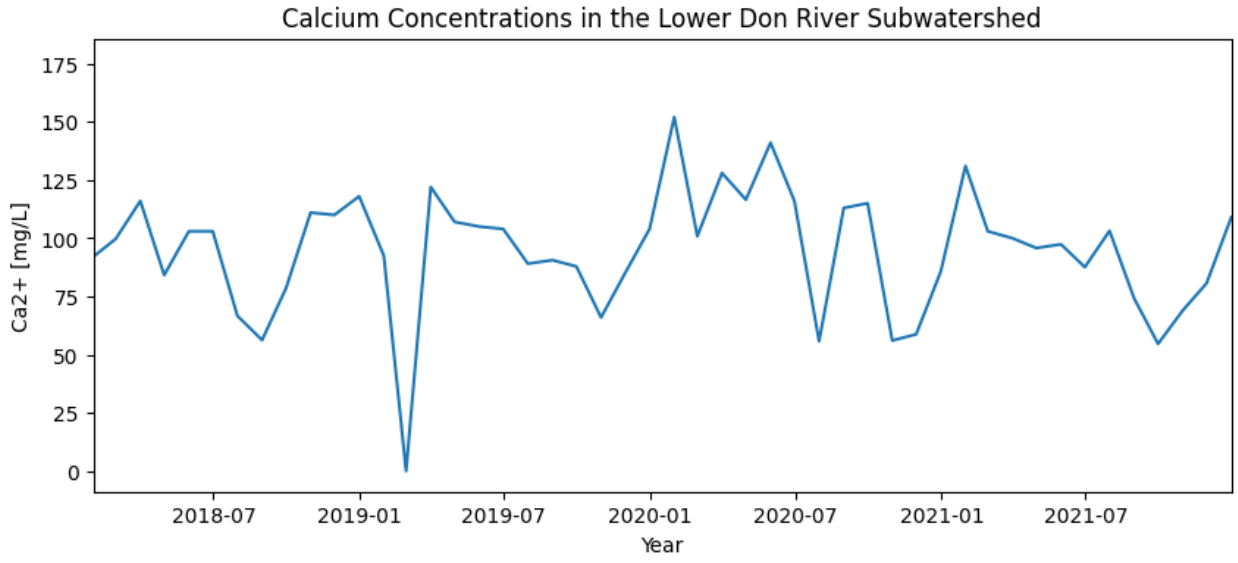


Figure 59 Observed Calcium concentrations in the Lower Don subwatershed.

These time series were synchronized, cleansed, and transformed before they were fed into the selected deep learning model. In the first approach, the number of observations

(rows) was 2052 and 56 features (columns) after preprocessing, whereas in the second approach, there were 1368 observations and 112 features. One of the important issues that had to be addressed arose from missing values.

4.1.2 Imputing Missing Data

Imputing missing values is the process of estimating or filling in the missing data points in a dataset with values that are predicted or derived based on the available information. This is often done to ensure the continuity and integrity of the dataset for further analysis, modelling, or visualization. In the context of the data collected from all the stations, missing values were prevalent (Figure 60). For historical data (1964 – 1999), erroneous values were present due to manual transcription errors. Detection limits and details of analytical methods for some historical data were unavailable, including missing or undefined value qualifiers and remark codes. Different imputation methods were considered in this study.

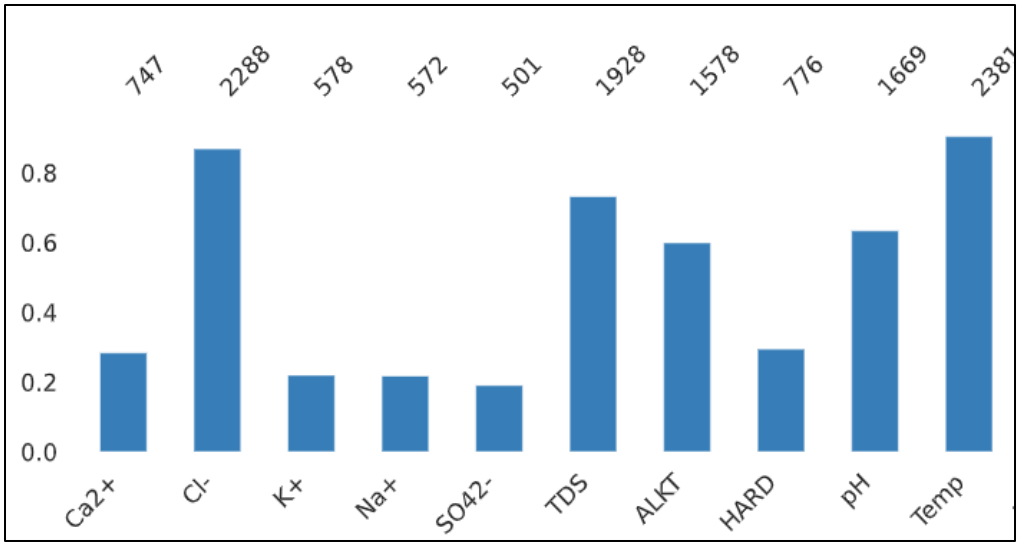


Figure 60 Percentage of Missing Values of Selected Water Quality Parameters

Imputation methods were examined, including Mean/Median Imputer, Gaussian End Tail Imputer, K-Nearest Neighbor Imputer, and Iterative Imputer. Since the data was organized on a monthly time scale, the initial imputation technique utilized each parameter's available mean monthly observations to populate the missing months accordingly. In Figure 61, a conspicuous pattern between 2000 and 2015 shows similar amplitudes and wavelengths. Duplicate values characterize this uniform periodicity, potentially impeding the model's ability to discern distinct patterns from historical data and consequently affecting its predictive capabilities for future values.

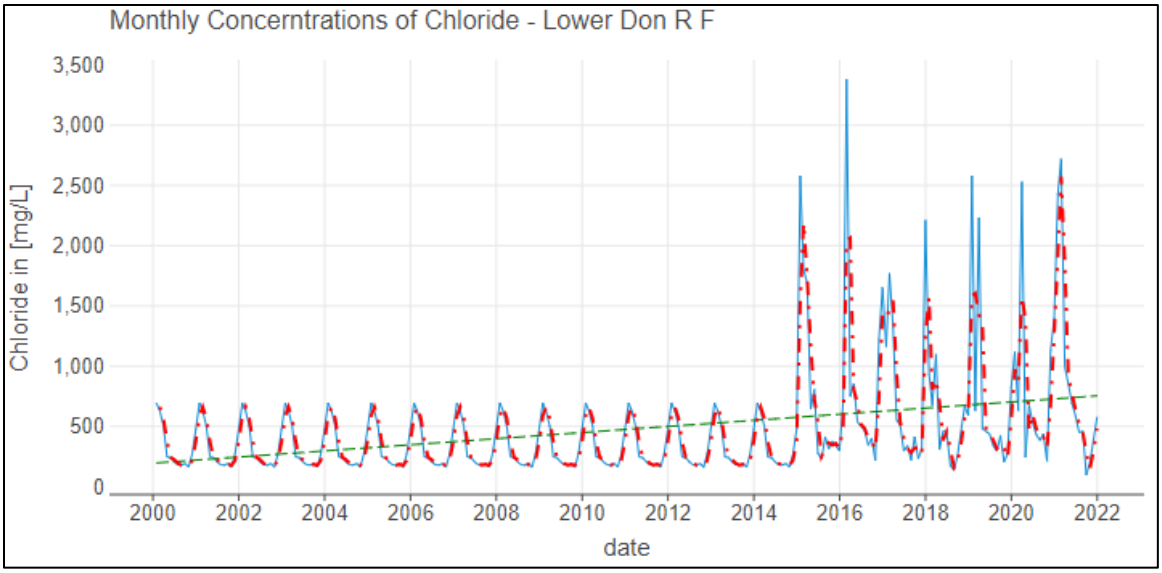


Figure 61 Missing values imputed using monthly means.

Additional imputation techniques were subjected to analysis. The dataset's missing value percentage was computed to determine the most suitable approach, revealing it to be around 50%. Following this, a subset of the dataset from 2015 to 2021, featuring a significant quantity of non-missing values (less than 1% missing data), was selected and treated as a complete dataset. By matching the absent value percentage from the original dataset, an equivalent number of values were removed from this sampled dataset.

This enabled the creation of an ensemble of machine learning algorithms, such as the Random Forest Regressor, to systematically evaluate various imputation techniques using a time series split methodology (Figure 64). The dataset was divided into five folds for this purpose. The performance scores of each technique are visually depicted in Figure 62. Remarkably, the iterative imputation method emerged as the most suitable technique in terms of minimal error (excluding the full dataset), boasting an MSE of 0.058 and a standard deviation of 0.070.

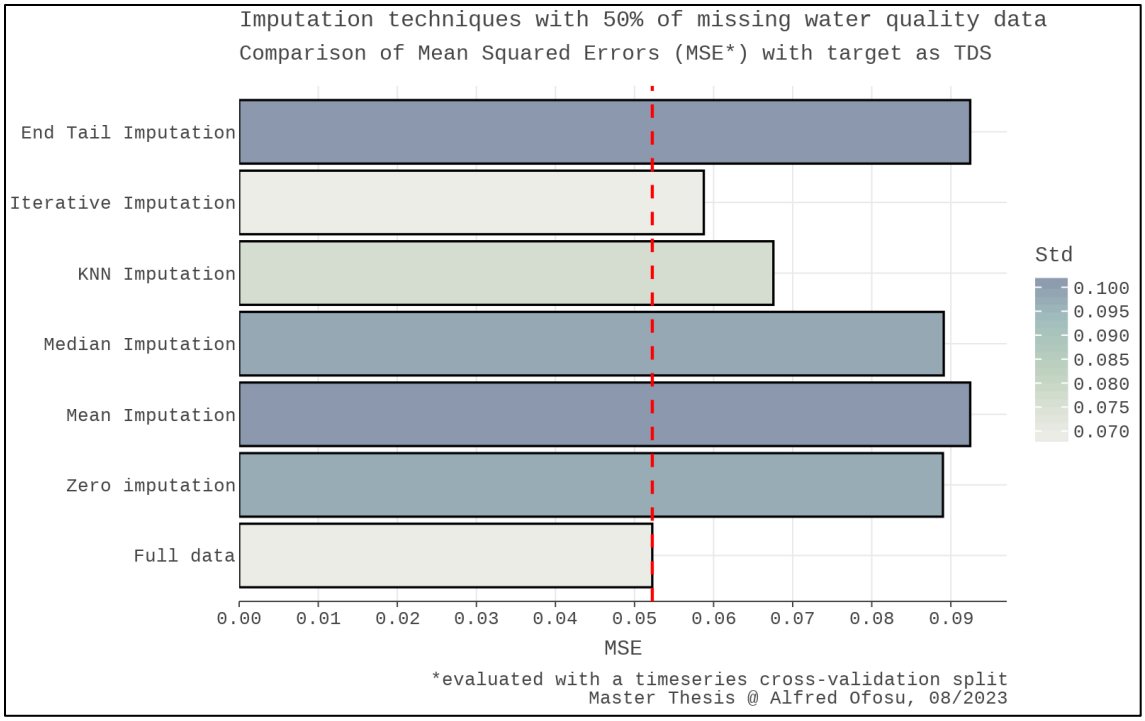


Figure 62 Results of Different Imputation Techniques

Figure 63 illustrates a comparison among the three alternative imputation methods, highlighting the notably superior performance of the iterative imputer in handling missing values. The iterative imputer was able to impute the seasonality and dynamics between 1995 and 2000 showing a better approximation than the others. The iterative imputer employs a distinctive approach by modelling each feature with absent data as a function of the remaining features, employing this estimation for imputation purposes. This process unfolds through an iterative cycle wherein, during each step, a specific feature column serves as the output variable y, while the remaining feature columns function as input variables X. A regressor is trained on the known y-values with their corresponding X-values. Subsequently, the regressor predicts the absent y-values. This iterative methodology is successively applied to each feature, repeating for a predetermined number of imputation rounds as specified by the model parameter **"max_iter"** (Pedregosa et al., 2011)

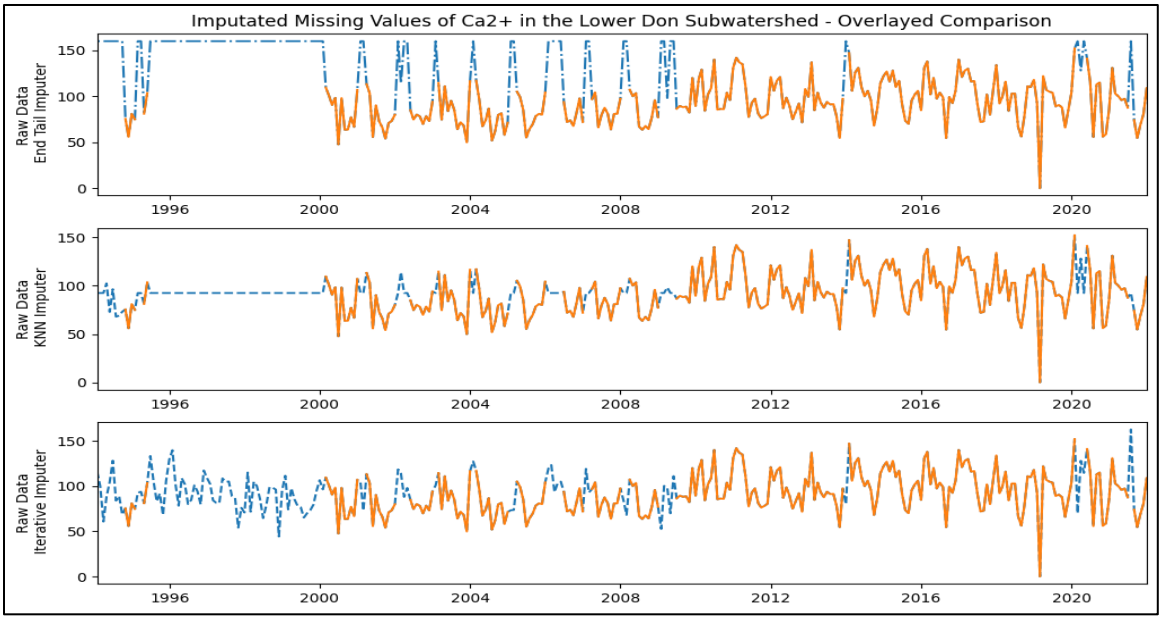


Figure 63 Comparing imputation methods (End Tail Imputer vs. KNN Imputer vs. Iterative Imputer)

It's important to highlight that the chosen water quality parameters, excluding temperature values, predominantly comprise positive values and to accommodate this characteristic, it is imperative to explicitly configure this in the arguments of the iterative imputer. Specifically, the parameter **"min_value"** must be specified as 0.0, ensuring that imputed values align with the intrinsic nature of these parameters.

Moreover, a comparative analysis of the parameter statistics before and after imputation was conducted to ascertain that the introduction of imputed values did not lead to an excessive level of noise within the dataset.

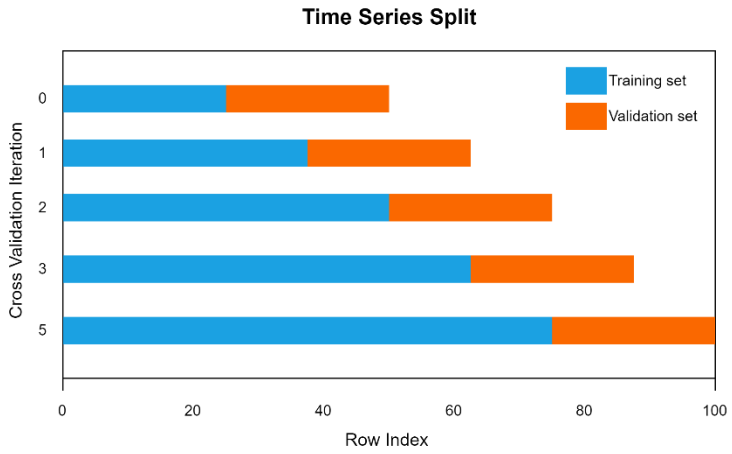


Figure 64 Time series split for cross-validation.

4.1.3 Feature Selection and Engineering

Feature engineering refers to selecting, modifying, or creating new features from raw data that machine learning algorithms will use as input variables. It involves transforming the raw data into a more suitable format, allowing the machine learning algorithm to learn patterns and make accurate predictions effectively.

Feature engineering is especially useful when observations or measurements return aggregated results. In the context of this study, concentrations of Carbonates and Bicarbonate ions were not determined analytically. However, their values can be calculated based on concentrations of other water quality parameters using equations 2 and 3 respectively. This allowed the inclusion of concentrations of these ions into a set of model features. Likewise, the water quality index for each month was incorporated using the water quality index (WQI) calculations from the Canadian Council of Ministers of the Environment (CCME). Since 2001, the CCME WQI has been used extensively in Canada and worldwide to report water quality. The CCME WQI is denoted as

$$CCMEWQI = 100 - \left(\frac{\sqrt{{F_1}^2 + {F_2}^2 + {F_3}^2}}{1.732} \right) \quad \textit{Equation 9}$$

where F_1 is the **Scope** and it represents the percentage of parameters that do not meet the guidelines at least once during the time frame under consideration (failed parameters) relative to the total number of parameters measured:

$$F_1 = \left(\frac{\text{Number of failed parameters}}{\text{Total number of parameters}} \right) \times 100 \quad \textit{Equation 10}$$

F_2 is the **Frequency** and it represents the percentage of individual test that do not meet the guidelines ("failed test"):

$$F_2 = \left(\frac{\text{Number of failed test}}{\text{Total number of test}} \right) \times 100 \quad \textit{Equation 11}$$

F_3 is the Amplitude and it represents the amount by which failed test do not meet their guidelines. F_3 is calculated in three further steps (CCME, 2017).

The number of times by which an individual concentration is greater than (or less than, when the objective minimum) the objective is termed an excursion and is denoted as:

test value < Objective $$excursion_i = \left(\frac{FailedTestValue_1}{Objective_1} \right) - 1 \quad \textit{Equation 12}$$

test value > Objective $$excursion_i = \left(\frac{\text{Objective}_1}{\text{FailedTestValue}_1} \right) - 1 \quad \textit{Equation 13}$$

$$\text{nse} = \frac{\sum_{i=1}^{n} excursion_i}{\#\ of\ tests} \quad \textit{Equation 14}$$

$$F_3 = \left(\frac{\text{nse}}{0.01\text{nse} + 0.01}\right) \qquad \textit{Equation 15}$$

where *nse* is the normalized sum of the excursion from objectives.

As stated in (CCME, 2017), the Water Quality Index (WQI) calculation requires a minimum of 8 parameters. Additionally, the WQI calculation necessitates a minimum of 4 parameters and four samples in a year to be valid. These considerations were carefully accounted for in this study. Consequently, essential physical water quality parameters, including pH, were chosen. This selection was made to ensure the availability of adequate parameters for the monthly WQI calculation for the Don River.

The WQI categories were transformed using ordinal encoders before their incorporation into each model.

Additionally, the models incorporated missing indicators, which were introduced into the dataset during the missing value imputations. This inclusion was intended to ensure the model was aware of missing values within the dataset.

4.1.4 Splitting Data for Training and Testing

Many hyperparameter optimization techniques cannot optimize cross-validation splits, especially in the context of deep learning models. Consequently, cross-validation splits were not considered during the hyperparameter tuning phase. In its place, a chronological data division strategy was employed, utilizing the `tf.keras.utils.timeseries_dataset_from_array()` utility from Keras/TensorFlow. This approach is particularly suited to handling time-series data, aligning well with the sequential nature commonly seen in deep learning tasks.

4.1.5 Data Scaling

Different scaling and transformation algorithms were assessed, including the Standard Scaler, Min-Max Scaler, Robust Scaler, Power Transformer, and Quantile Transformer. Among these options, the *Standard Scaler* was chosen not to suppress outliers as it was of interest to create a model that could also predict extreme values. Outliers are data points that significantly deviate from the main distribution, potentially affecting the overall shape of the distribution.

In instances where outlier investigation is crucial and extreme concentrations of water quality parameters need to be identified, it becomes necessary to scale the data while retaining its inherent characteristics. The Standard Scaler strikes a balance between scaling the data and preserving its original attributes, making it a suitable choice for this specific scenario.

.

The Standard Scaler is denoted as:

$$x_1 = \frac{x_i^{org} - \mu_i^{org}}{\sigma_i^{org}} \qquad \textit{Equation 16}$$

Where μ_i^{org} and σ_i^{org} are the mean and standard deviation of the original x variable at index i.

4.2 Hyperparameter Optimization

Hyperparameter optimization (HO), also known as hyperparameter tuning, refers to finding the optimal values for the hyperparameters of an ML model. Hyperparameters are the model's parameters defined in the model's algorithm. They determine the model architecture and specifics of model training. There are many hyperparameters, such as the number of hidden layers, neurons in each layer, the learning rate, batch size, and

regularization rates. HO is essential because the model's performance significantly relies on the values of these hyperparameters. Manually selecting these hyperparameters can be both costly and time-consuming. Fortunately, numerous Python libraries are readily available to automate the process of tuning hyperparameters within a defined search space. AutoML, Scikit, Ray Tune, Keras Tuner, Optuna, etc., are very popular among researchers. All these libraries provide baseline optimizers such as Grid Search, Random Search Bayesian Optimization, and additional proprietary optimizers. In the context of this research, Keras Tuner and Optuna were evaluated. The choice fell to Optuna because it was reliable and provided better outcomes than the Keras Tuner.

4.2.1 Loss Functions

The problem at hand must dictate the selection of a loss function. This study considered two loss functions- Mean Absolute Error (MAE) and Mean Squared Error (MSE) – for the following reasons. Water quality assessment relies on two important characteristics: instantaneous concentrations and the amount of matter passing through a cross-section of a water body over a given period. MSE allows to determine the most appropriate approximations of data over a given period, while the MAE takes into account the extreme values of concentrations.

MAE is denoted as

$$\mathrm{MAE} = \left(\frac{\sum_{i=1}^{n} |y_{pred_i} - y_i|}{\mathrm{n}}\right) \qquad \textit{Equation 17}$$

where y_{pred_i} is the predicted value, y_i the true value and n the total number of data points. MAE serves as a metric that gauges the total sum of errors between the actual and predicted values, and this sum is then divided by the number of samples in each batch size. In this context, the term "batch size" pertains to the number of training inputs

considered during each iteration of the process. MAE is less sensitive to outliers since it considers the absolute difference between the actual and predicted values. MAE is beneficial in understanding the magnitude of water quality concentrations in each predicted month.

MSE is denoted as

$$\text{MSE} = \left(\frac{\sum_{i=1}^{n}(y_i - y_{pred_i})^2}{\text{n}}\right) \qquad \textit{Equation 18}$$

MSE measures the average of the squares of the errors i.e., the average squared difference between the predicted and actual values. MSE is very sensitive to outliers because larger weights are given to larger errors and thus, in the context of this research, it serves the purpose of understanding the effects of outliers on the models.

Seven major ions were evaluated on each loss function for each approach making it in total 7 x 2 x 2 evaluations.

4.2.2 Optimization Algorithms

An optimization algorithm is required to adjust the weights of an ANN. In weight and bias optimization, the goal is to find the optimal values that minimize the model's loss function. Various optimization algorithms are available for this purpose, including Stochastic Gradient Descent (SGD), Root Mean Square Propagation (RMSProp), and Adam (Adaptive Moment Estimation)

Adam was chosen for this research as the optimization algorithm responsible for updating the weights and biases of a deep learning model after each iteration. Adam is selected for this research due to its advantages over the other two optimizers. SGD can encounter the vanishing gradient problem, which affects its convergence speed and stability during training. RMSProp, on the other hand, adjusts the learning rates for each

parameter individually, which can lead to biased updates and slow convergence in some cases.

Adam combines the benefits of both SGD and RMSProp. It maintains separate learning rates for each parameter like RMSProp, but also incorporates momentum from past gradients to ensure more stable updates. This adaptive learning rate adjustment, along with the momentum, helps Adam converge faster and more consistently in various scenarios. Therefore, considering its improved convergence properties and ability to handle different types of data, Adam is a suitable choice for optimizing the weights and biases of the developed water quality models.

Adam combines the benefits of both SGD and RMSProp. It maintains separate learning rates for each parameter like RMSProp, but also incorporates momentum from past gradients to ensure more stable updates. This adaptive learning rate adjustment, along with the momentum, helps Adam converge faster and more consistently in various scenarios. Therefore, considering its improved convergence properties and ability to handle different types of data, Adam is a suitable choice for optimizing the weights and biases of the deep learning models in this research.

4.2.3 Generalization Techniques

Generalizing a model to a test dataset, which the model hasn't encountered before, is achieved when the model can make predictions with minimal variance and bias. In the context of supervised regression tasks, a good validation loss doesn't guarantee optimal performance, as the model could still suffer from either overfitting or underfitting.

Overfitting occurs when a model performs well on the training dataset but poorly on the validation dataset or new data. This indicates that the model has learned the training data's noise and specific details to an extent that it can't effectively generalize to new

examples. Underfitting is the opposite, and it occurs when there is still room for improvement on the train dataset.

To overcome these caveats, (Hinton et al., 2012) developed a way to control overfitting called Dropouts. Dropout, a widely used regularization method, is particularly effective in large models. It involves adjusting the weights of the incoming connections to hidden units, enabling the model to learn feature detectors that facilitate accurate output predictions when presented with input vectors.

During each training iteration, every hidden unit is randomly excluded from the network with a probability of 0.5. This strategy prevents a hidden unit from depending on the presence of other hidden units. Dropout can also be perceived as an efficient method for performing model averaging with neural networks, enhancing the network's robustness and ability to generalize (Hinton et al., 2012).

4.3 Modelling Results

In both research approaches, the HO process delved into expansive search spaces for each model architecture. This exploration encompassed settings such as a maximum of 5 layers and hidden layer units varying between 32 and 1024. The study also considered five distinct activation functions, relu, elu, selu, gelu, and tanh, as part of the optimization process. Additionally, the option of setting a dropout rate was included, right before the output layer. The choices made in this first approach were very ambiguous to understand the impact a larger search space would have on the validation dataset.

It is imperative to note that HO makes decisions based on signals computed from the validation dataset, a process that can inadvertently lead to overfitting, as emphasized by Chollet (Chollet, 2022). Consequently, much caution was taken to not optimize hyperparameters such as SGD which involved tunning additional model parameters such as their momentum and learning rate.

The HO process was structured to initially fine-tune the models. Subsequently, the optimal hyperparameters obtained were employed to retrain the model on both the training and validation sets, with the intention of determining the most suitable epoch. To prevent unnecessary prolonged training, a Callback mechanism was incorporated, specifically with a patience of 5 epochs, which would halt the training if no improvement was observed. The hyperparameters were manually tuned if results showed (1) overfitting or (2) poor errors.

4.3.1 Research Approach 1

The first approach is grounded on the assumption that significant dependencies in major ion concentrations from the upstream and downstream sections of a river should be reflected in the available data sets. Consequently, the entire watershed system was treated as a singular entity to predict seven major ion concentrations in each cross-section of the Don River. In this context, the data collected from all monitoring stations (water quality, hydrometric, and meteorological stations) was aggregated and employed in the modelling process for each water quality parameter. Furthermore, one-hot-shot labels indicated whether a specific water quality parameter failed (1) or not (0) based on the CCMEWQI guidelines, missing-value indicators, and water quality indices. The primary objective revolved around forecasting concentrations for a future month (t+1) based on historical data leading up to that point.

Research Approach 1 will exclusively showcase outcomes obtained using the Loss function MAE to streamline the presentation of model results. At the same time, Approach 2 will exclusively feature results derived from the Loss function MSE.

Table 2 presents the computational results obtained by employing data from all monitoring stations within the Don River watershed. Across the range of models investigated, the DNN architecture accurately predicted most water quality parameters,

except for Magnesium concentrations. On the contrary, CNN's performance was subpar. Convolutional Neural Networks (CNNs) are primarily tailored for image-based predictions and didn't fare well in this context.

Table 2 MAE and MSE modelling results in research approach 1.

Major Ion	sLNN		DNN		CNN		RNN	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Calcium	0.2701	0.1158	**0.1646**	**0.0431**	0.4748	0.3859	0.3889	0.2872
Chloride	0.3201	0.2071	**0.1855**	**0.1804**	0.5343	0.5719	0.2943	0.2984
Potassium	**0.1250**	0.0919	0.2115	**0.0849**	0.6223	0.6240	0.3255	0.2014
Magnesium	0.3087	**0.0732**	0.2138	0.1184	0.5018	0.3448	0.2082	0.1524
Sodium	0.4310	0.2429	**0.3102**	**0.1146**	0.5135	0.5311	0.3146	0.2071
Sulphate	0.2753	0.3098	**0.1494**	**0.1228**	0.6975	1.0054	0.3052	0.3059
TDS	0.4307	0.2671	**0.2717**	**0.1390**	0.6549	0.7285	0.3295	0.3146

The outcomes of the calculations pertaining to the baseline model (sLNN) are outlined below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset. The actual values are depicted by the black line, while the predicted values are illustrated by the red dashed lines. Both the baseline model and the DNN exhibited superior performance compared to the CNN and RNN. This suggests that employing intricate models is unnecessary for accurately predicting water quality concentrations in this approach.

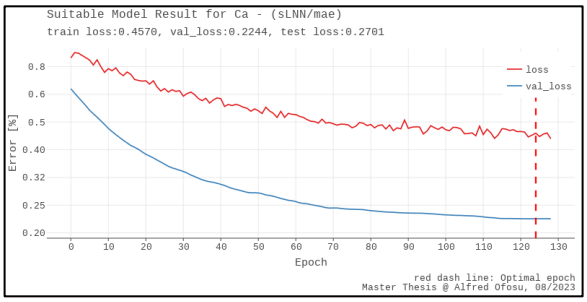


Figure 65 Modelling result for Calcium - sLNN

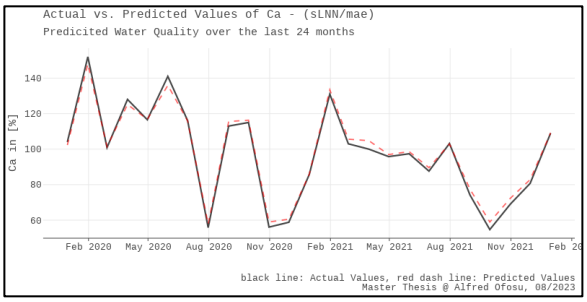


Figure 66 Predicted concentrations of Calcium - sLNN

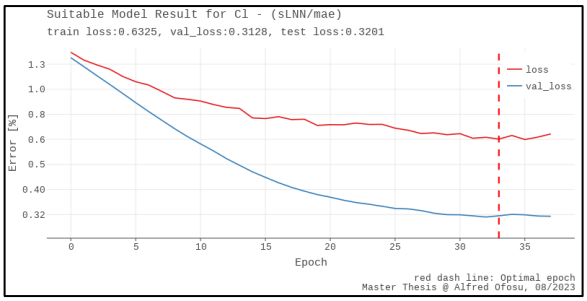


Figure 67 Modelling results, Chloride - sLNN

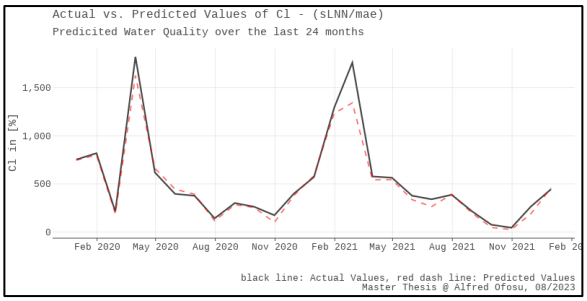


Figure 68 Predicted concentrations of Chloride - sLNN

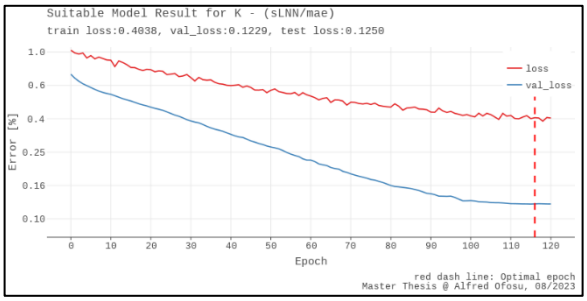


Figure 69 Modelling results, Potassium - sLNN

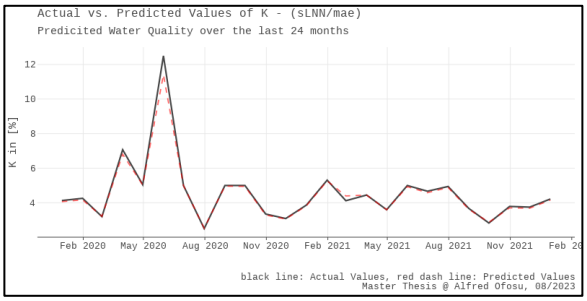


Figure 70 Predicted concentrations of Potassium - sLNN

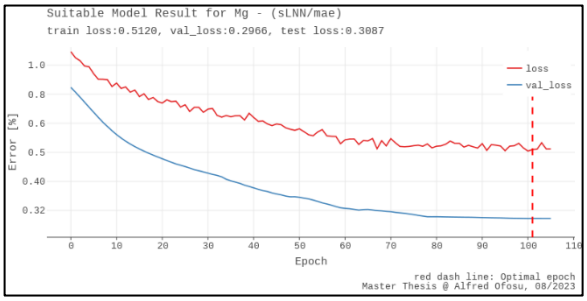


Figure 71 Modelling results, Magnesium - sLNN

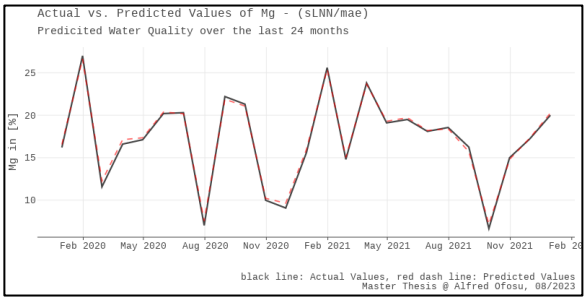


Figure 72 Predicted concentrations of Magnesium - sLNN

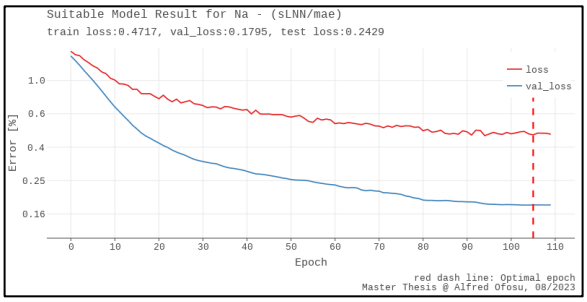


Figure 73 Modelling results, Sodium - sLNN

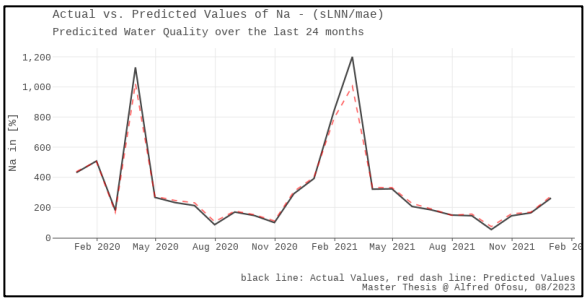


Figure 74 Predicted concentrations of Sodium - sLNN

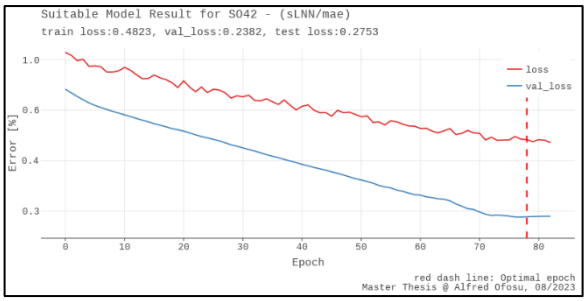


Figure 75 Modelling results, Sulphate - sLNN

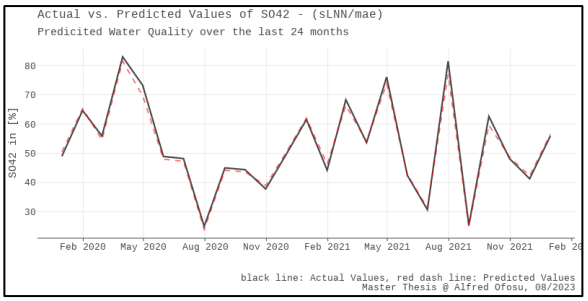


Figure 76 Predicted concentrations of Sulphate - sLNN

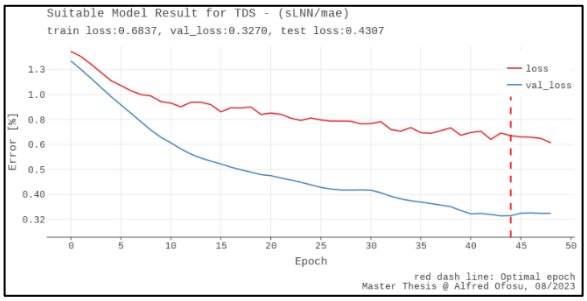


Figure 77 Modelling results, TDS - sLNN

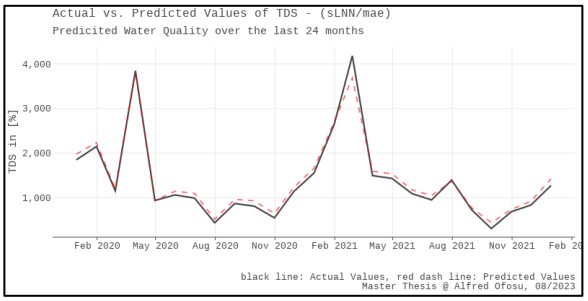


Figure 78 Predicted concentrations of TDS - sLNN

The results of the computations on the deep neural network (DNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

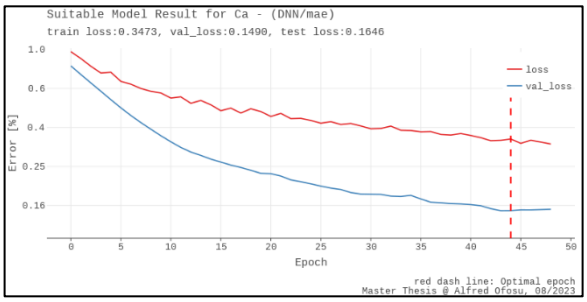


Figure 79 Modelling results, Calcium - DNN

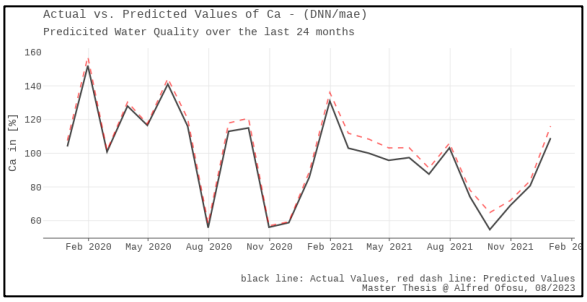


Figure 80 Predicted concentrations of Calcium- DNN

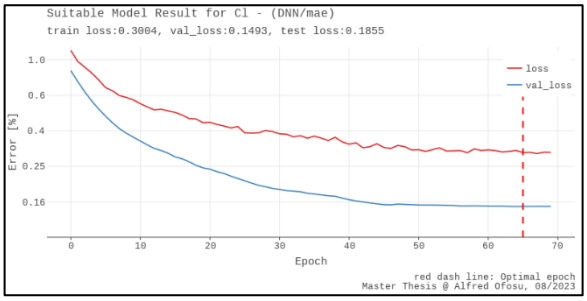


Figure 81 Modelling results, Chloride - DNN

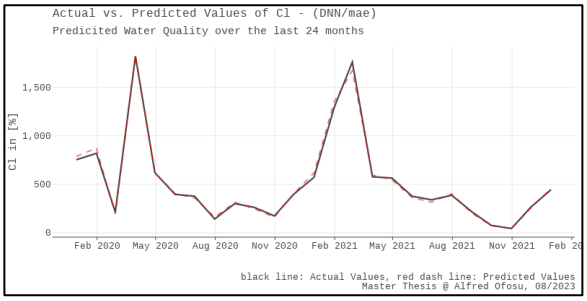


Figure 82 Predicted concentrations of Chloride - DNN

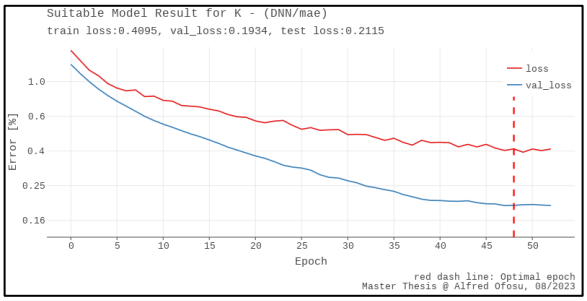


Figure 83 Modelling results, Potassium - DNN

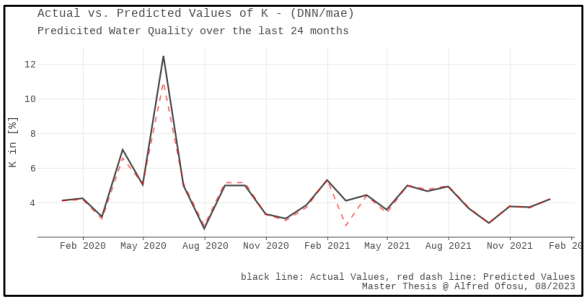


Figure 84 Predicted concentrations of Potassium - DNN

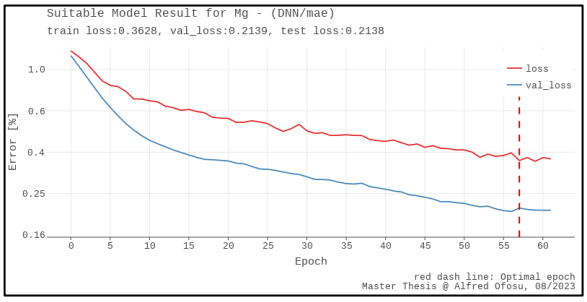


Figure 85 Modelling results, Magnesium - DNN

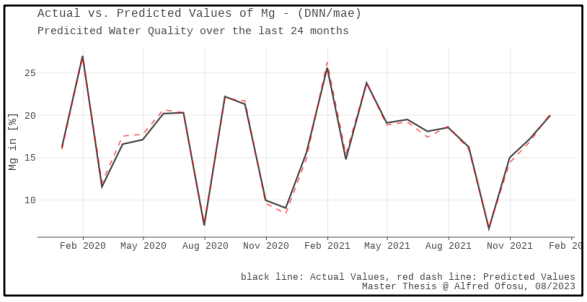


Figure 86 Predicted concentrations of Magnesium - DNN

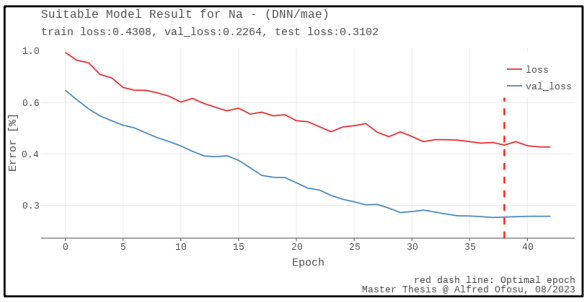


Figure 87 Modelling results, Sodium - DNN

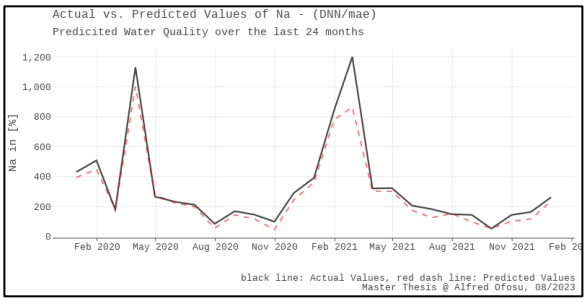


Figure 88 Predicted concentrations of Sodium - DNN

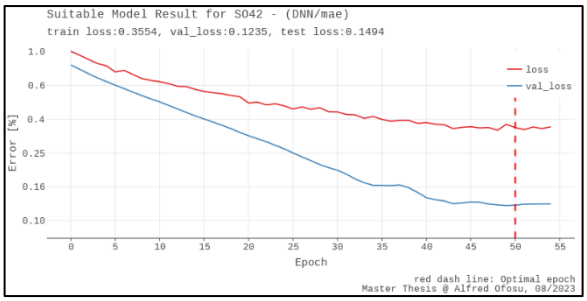


Figure 89 Modelling results, Sulphate - DNN

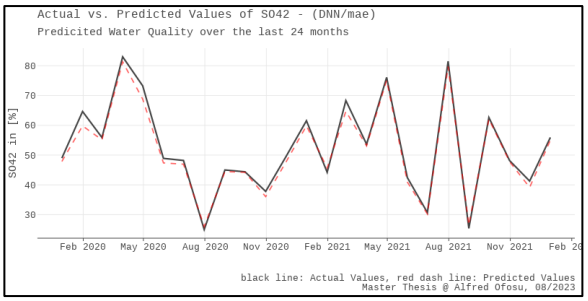


Figure 90 Predicted concentrations of Sulphate - DNN

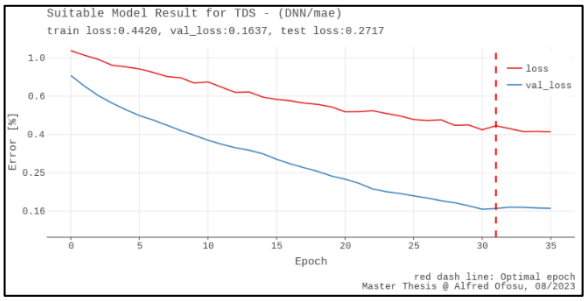


Figure 91 Modelling results, TDS - DNN

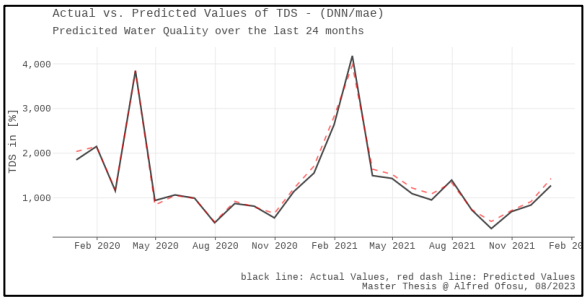


Figure 92 Predicted concentrations of TDS - DNN

The results of the computations on the convolutional neural network (CNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

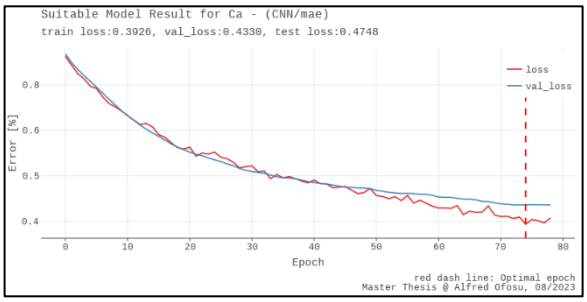


Figure 93 Modelling results, Calcium - CNN

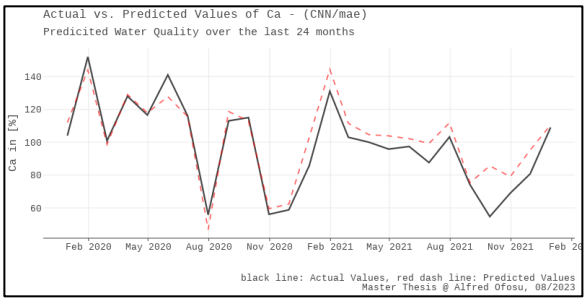


Figure 94 Predicted concentrations of Calcium - CNN

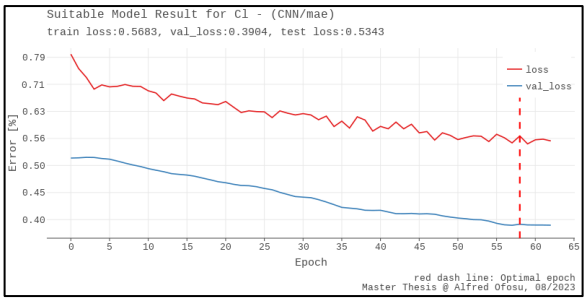


Figure 95 Modelling results, Chloride - CNN

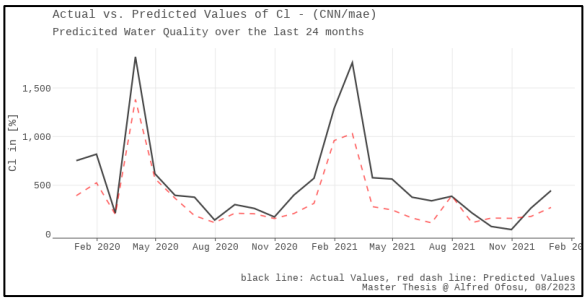


Figure 96 Predicted concentrations of Chloride - CNN

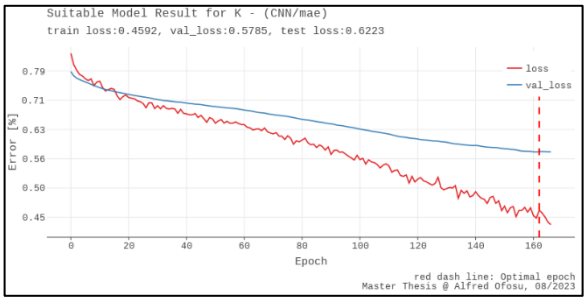


Figure 97 Modelling results, Potassium - CNN

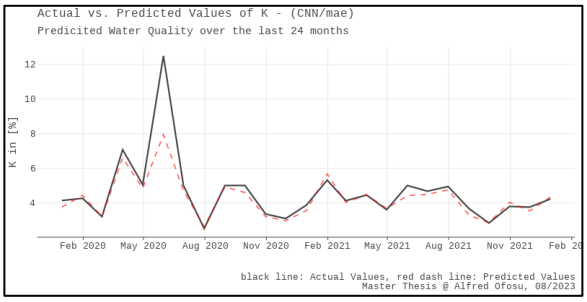


Figure 98 Predicted concentrations of Potassium - CNN

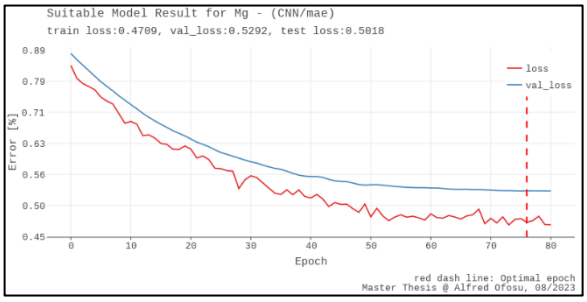


Figure 99 Modelling results, Magnesium - CNN

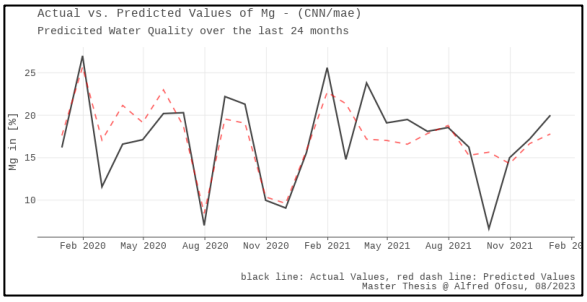


Figure 100 Predicted concentrations of Magnesium - CNN

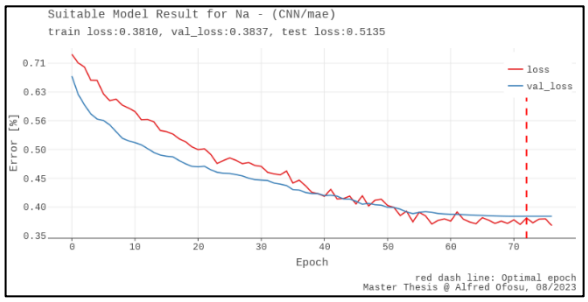


Figure 101 Modelling results, Sodium - CNN

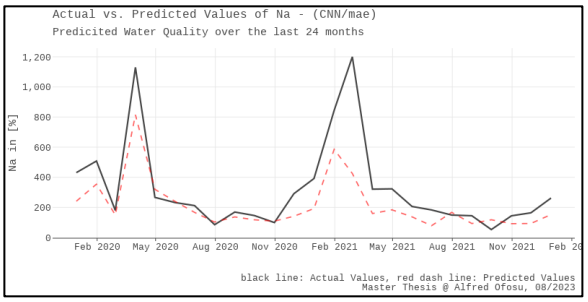


Figure 102 Predicted concentrations of Sodium - CNN

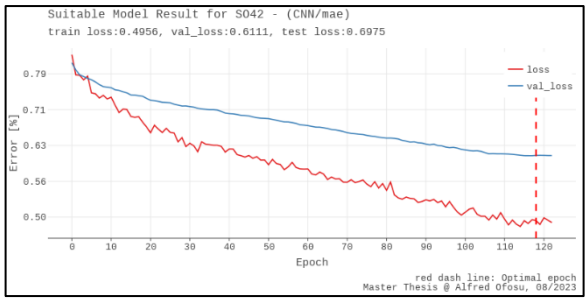


Figure 103 Modelling results, Sulphate - CNN

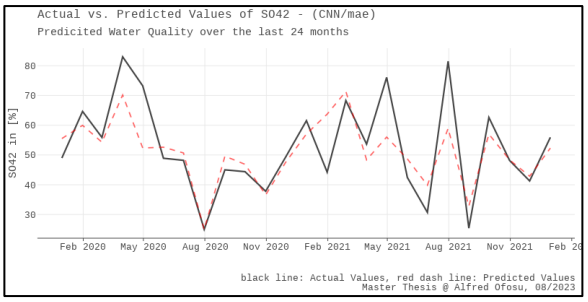


Figure 104 Predicted concentrations of Sulphate - CNN

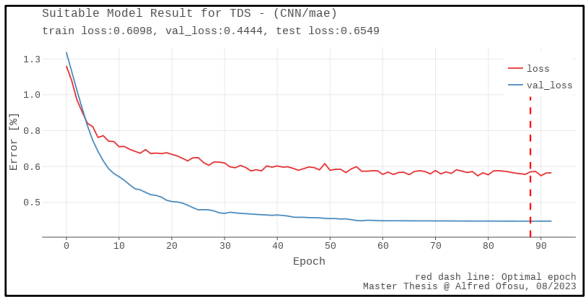


Figure 105 Modelling results, TDS - CNN

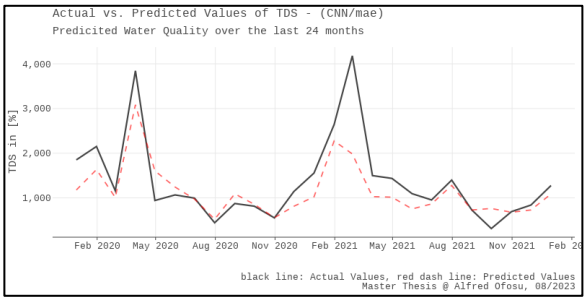


Figure 106 Predicted concentrations of TDS - CNN

The results of the computations on the recurrent neural network (RNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

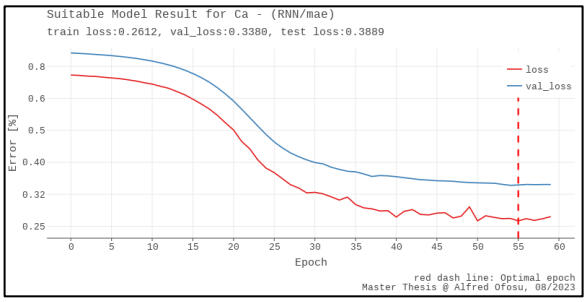


Figure 107 Modelling results, Calcium - RNN

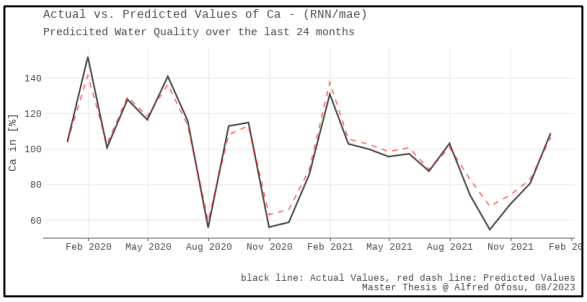


Figure 108 Predicted concentrations of Calcium - RNN

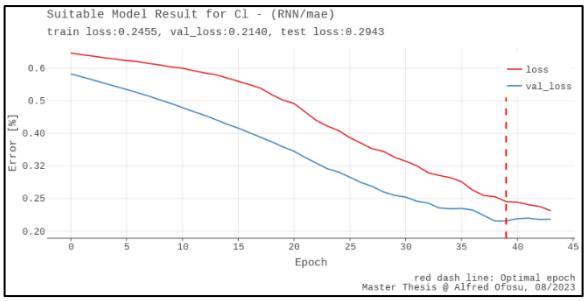


Figure 109 Modelling results, Chloride - RNN

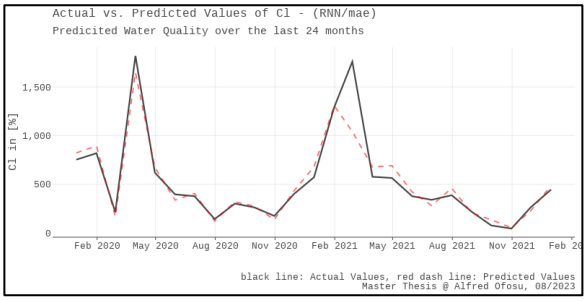


Figure 110 Predicted concentrations of Chloride - RNN

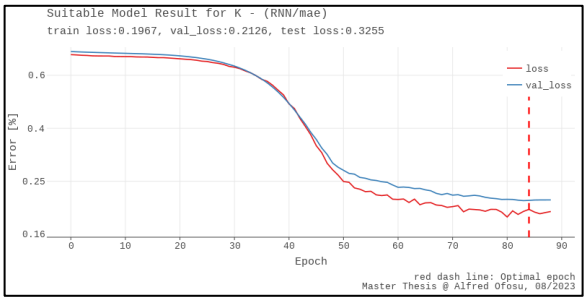


Figure 111 Modelling results, Potassium - RNN

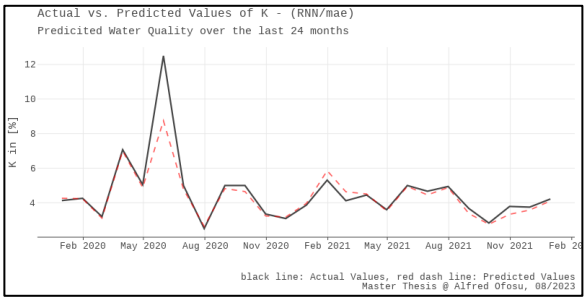


Figure 112 Predicted concentrations of Potassium - RNN

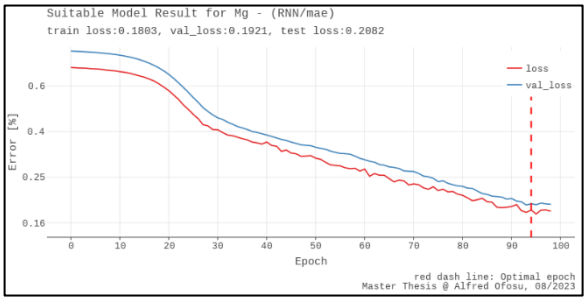


Figure 113 Modelling results, Magnesium - RNN

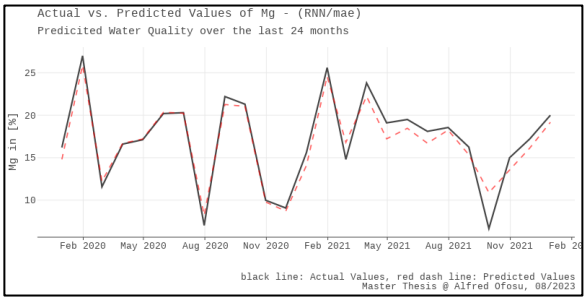


Figure 114 Predicted concentrations of Magnesium - RNN

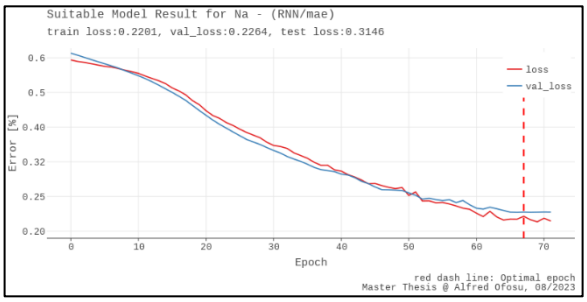


Figure 115 Modelling results, Sodium - RNN

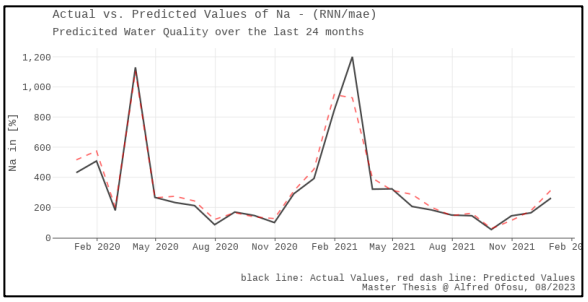


Figure 116 Predicted concentrations of Sodium - RNN

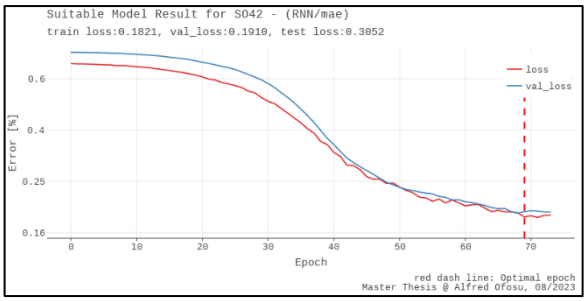


Figure 117 Modelling results, Sulphate - RNN

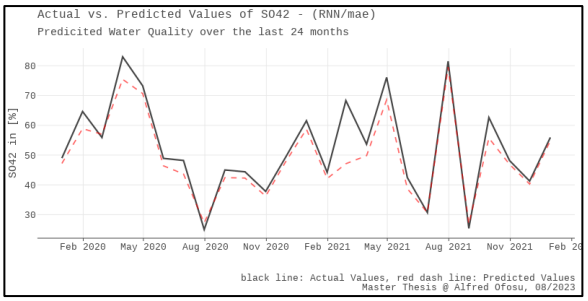


Figure 118 Predicted concentrations of Sulphate - RNN

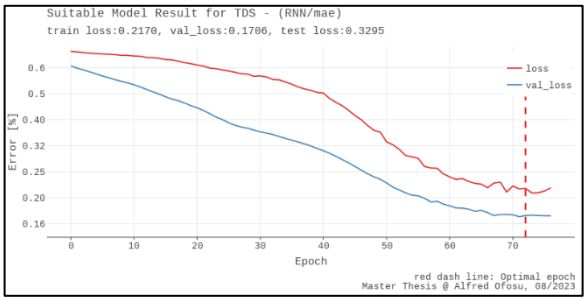


Figure 119 Modelling results, TDS - RNN

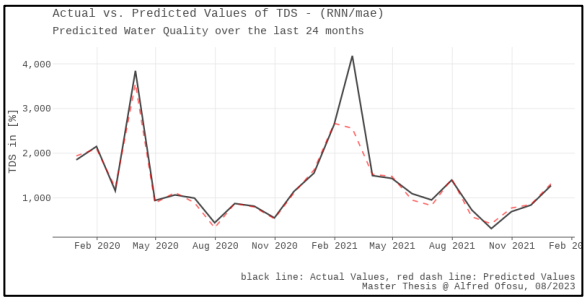


Figure 120 Predicted concentrations of TDS - RNN

4.3.2 Research Approach 2

The second approach posits that water quality from the upstream locations influences concentrations of the investigated water constituents in the downstream river cross-section, allowing upstream observations to forecast downstream water quality.

The input data for this approach were derived from monitoring stations in both the upstream rivers, namely the West Don River and the East Don River. On the other hand, the output target data were extracted specifically from the monitoring stations within the Lower Don River. This approach modelled the variations in water quality parameters across these distinct watershed sections, thereby providing a holistic understanding of the water quality dynamics within the entire Don River system. Furthermore, one-hot-shot labels indicated whether a specific water quality parameter failed (1) or not (0) based on the CCMEWQI guidelines and missing-value indicators.

In this approach, the RNN model exhibited precise predictions for concentrations of all water parameters, Table 3. Notably, since the lower stream parameters were forecasted using upper stream water quality features, certain elements, such as the CCMEWQI and failed guideline indicators, were excluded. This exclusion was because these elements were not formulated regarding the quality of the lower stream.

Table 3 MAE and MSE modelling results in research approach 2.

Major Ion	sLNN		DNN		CNN		RNN	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Calcium	1.1738	1.2997	0.8535	0.9087	0.7122	0.6921	**0.3001**	**0.2458**
Chloride	0.6925	1.5072	0.8592	1.5171	0.7471	1.6125	**0.2975**	**0.1997**
Potassium	0.8951	1.5006	0.8026	1.3342	0.7173	0.9109	**0.3186**	**0.1132**
Magnesium	0.8924	0.8262	0.7647	1.0006	0.7088	0.7636	**0.3448**	**0.1569**
Sodium	0.7596	1.2821	0.6961	1.4832	0.7353	1.4164	**0.2585**	**0.1333**
Sulphate	0.8461	1.9475	0.7606	1.0573	0.8737	0.9313	**0.3661**	**0.1406**
TDS	1.2588	1.9155	0.8284	1.2475	0.8813	1.2643	**0.7515**	**0.1273**

The results of the computations on the baseline (sLNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

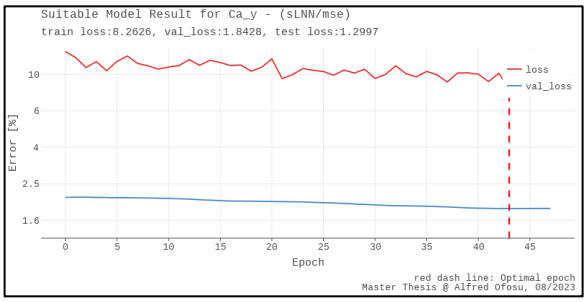


Figure 121 Modelling results, Calcium – sLNN, 2

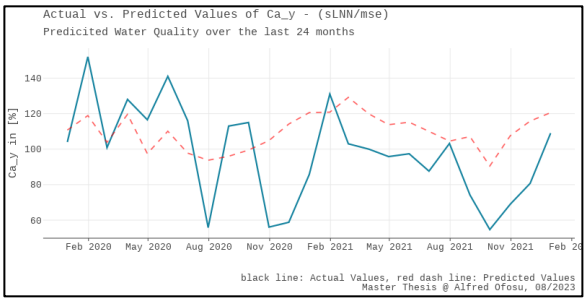


Figure 122 Predicted concentrations of Calcium – sLNN, 2

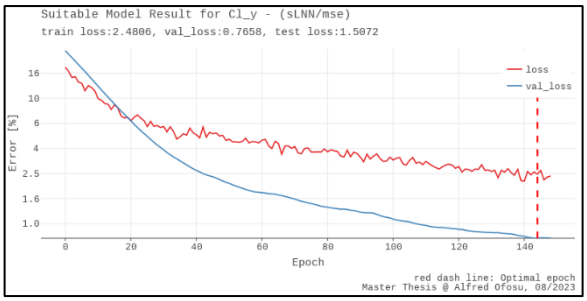


Figure 123 Modelling results, Chloride – sLNN, 2

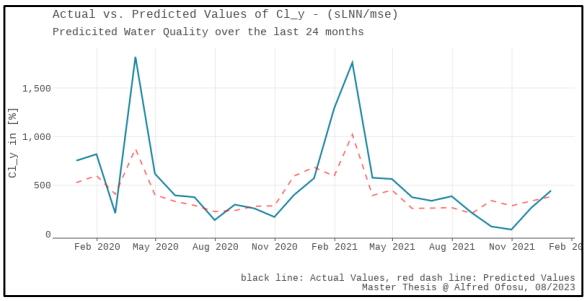


Figure 124 Predicted concentrations of Chloride – sLNN, 2

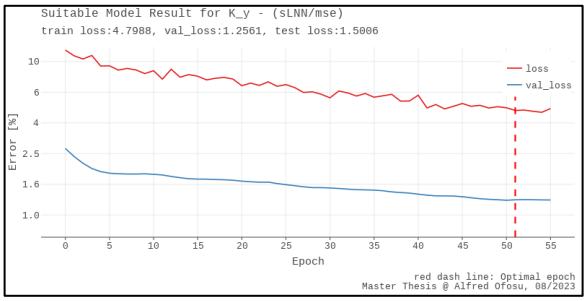


Figure 125 Modelling results, Potassium – sLNN, 2

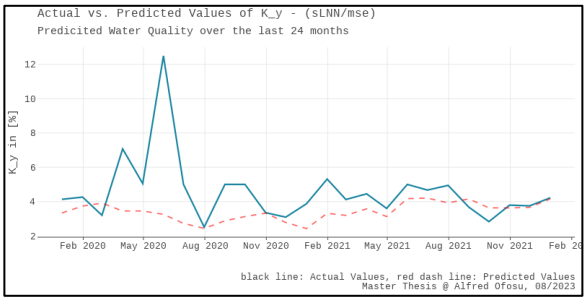


Figure 126 Predicted concentrations of Potassium – sLNN, 2

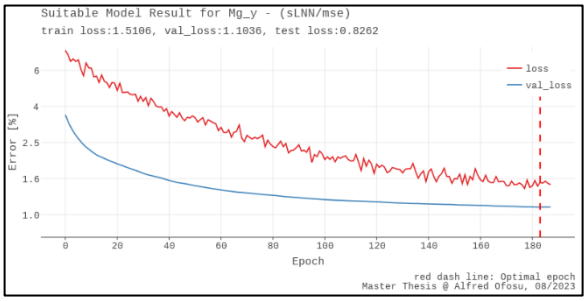


Figure 127 Modelling results, Magnesium – sLNN, 2

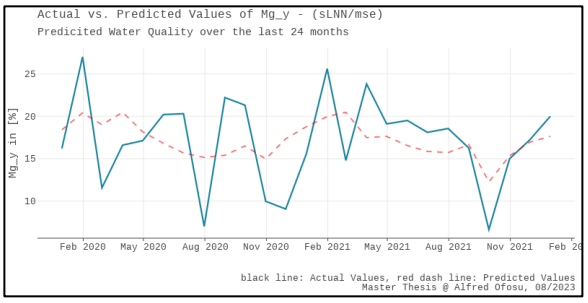


Figure 128 Predicted concentrations of Magnesium – sLNN, 2

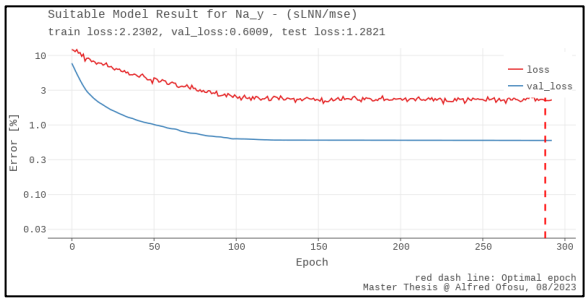


Figure 129 Modelling results, Sodium – sLNN, 2

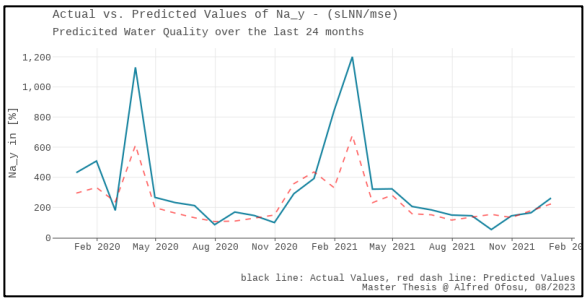


Figure 130 Predicted concentrations of Sodium – sLNN, 2

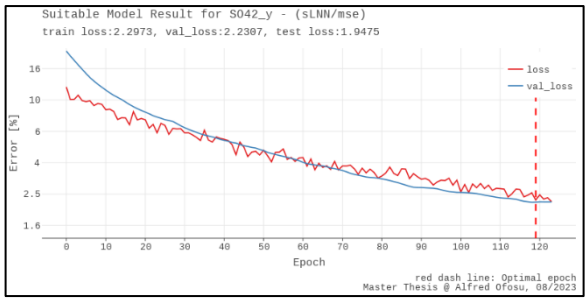


Figure 131 Modelling results, Sulphate – sLNN, 2

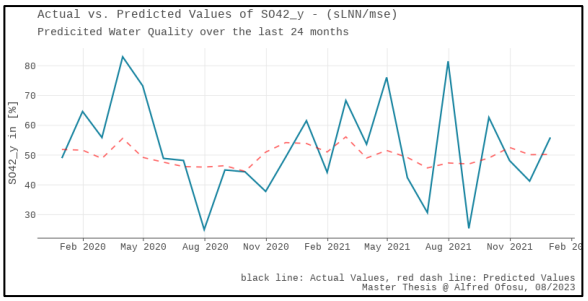


Figure 132 Predicted concentrations of Sulphate – sLNN, 2

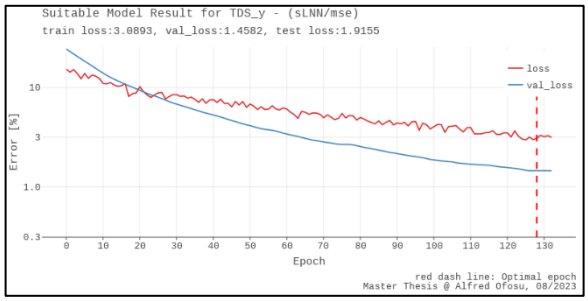


Figure 133 Modelling results, TDS – sLNN, 2

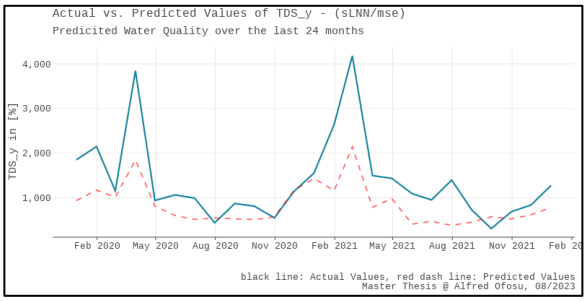


Figure 134 Predicted concentrations of TDS – sLNN, 2

The results of the computations on the deep neural network (DNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

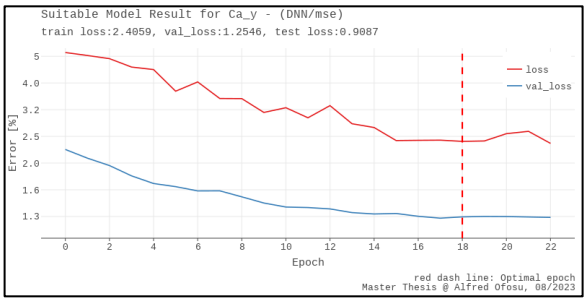


Figure 135 Modelling results, Calcium – DNN, 2

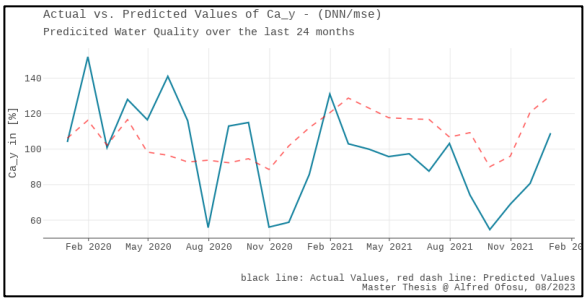


Figure 136 Predicted concentrations of Calcium – DNN, 2

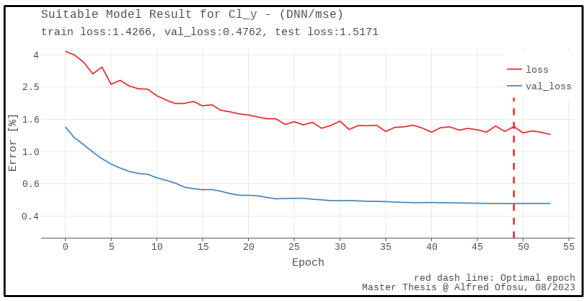


Figure 137 Modelling results, Chloride – DNN, 2

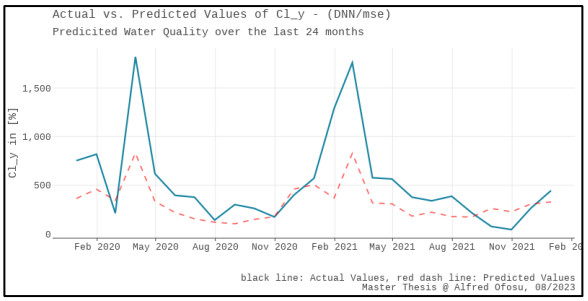


Figure 138 Predicted concentrations of Chloride – DNN, 2

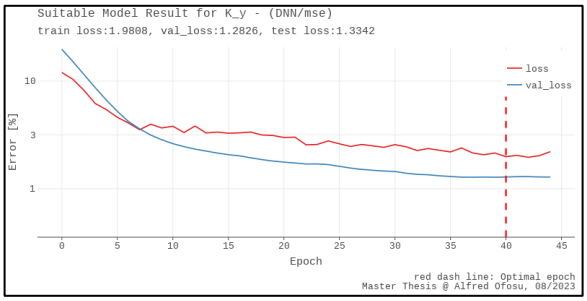


Figure 139 Modelling results, Potassium – DNN, 2

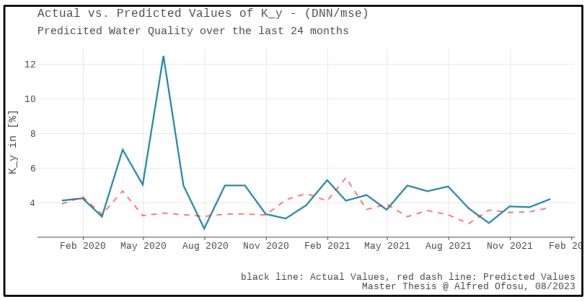


Figure 140 Predicted concentrations of Potassium – DNN, 2

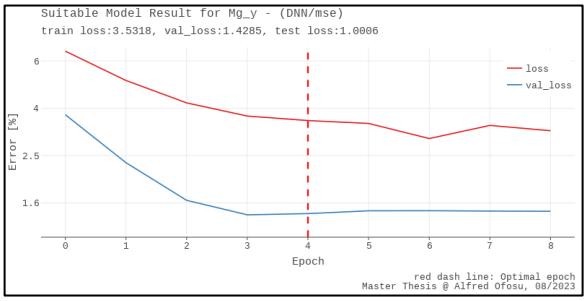


Figure 141 Modelling results, Magnesium – DNN, 2

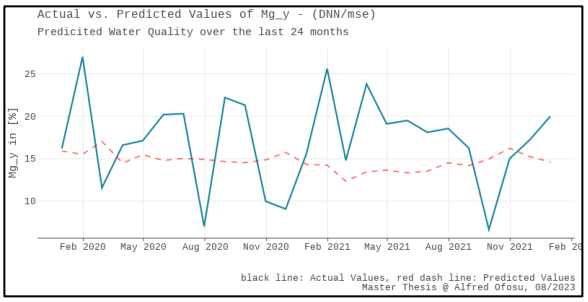


Figure 142 Predicted concentrations of Magnesium – DNN, 2

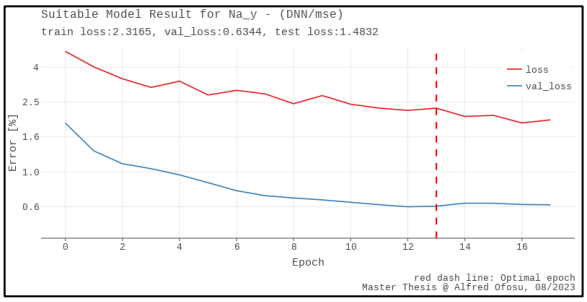


Figure 143 Modelling results, Sodium – sLNN, 2

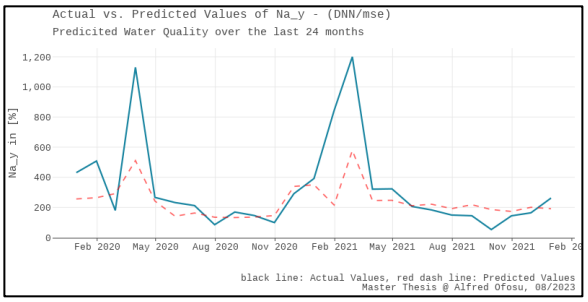


Figure 144 Predicted concentrations of Sodium – DNN, 2

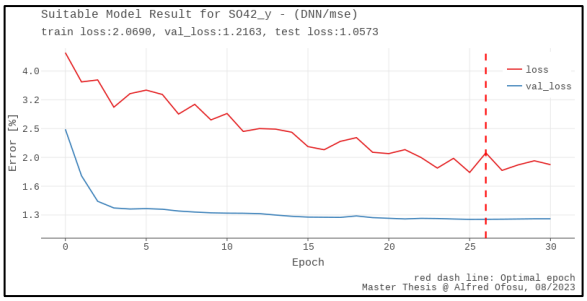


Figure 145 Modelling results, Sulphate – DNN, 2

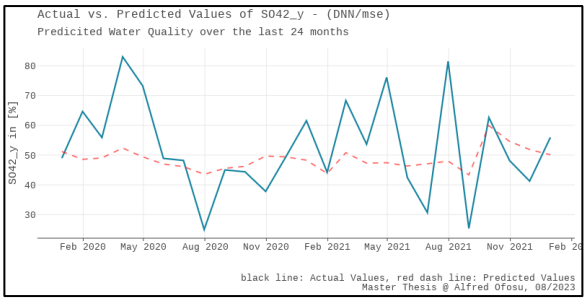


Figure 146 Predicted concentrations of Sulphate – DNN, 2

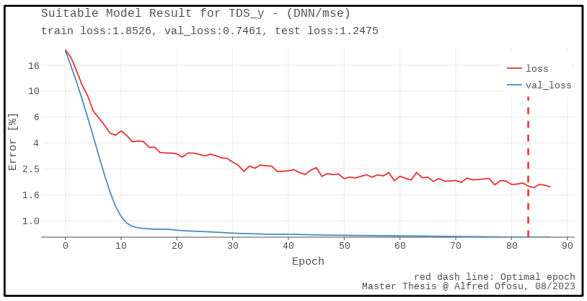


Figure 147 Modelling results, TDS – DNN, 2

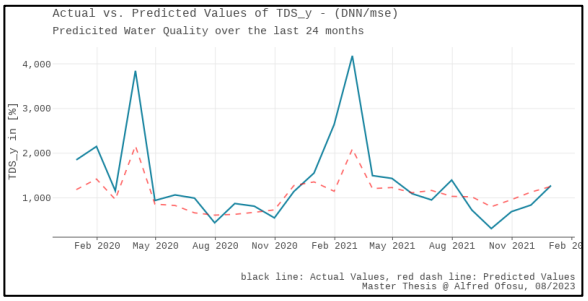


Figure 148 Predicted concentrations of TDS – DNN, 2

The results of the computations on the convolutional neural network (DNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

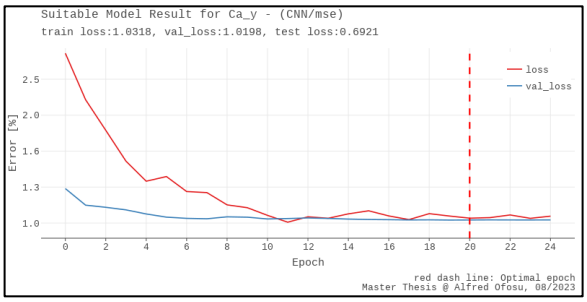


Figure 149 Modelling results, Calcium – sLNN, 2

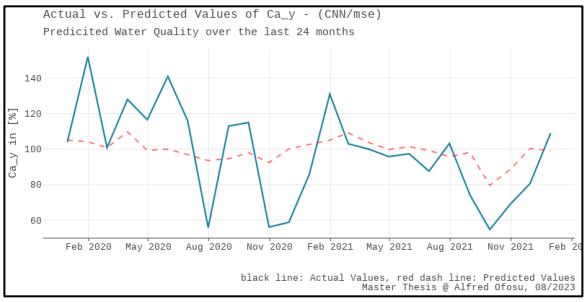


Figure 150 Predicted concentrations of Calcium – sLNN, 2

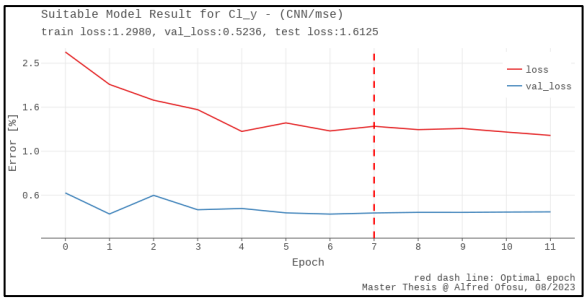


Figure 151 Modelling results, Chloride – sLNN, 2

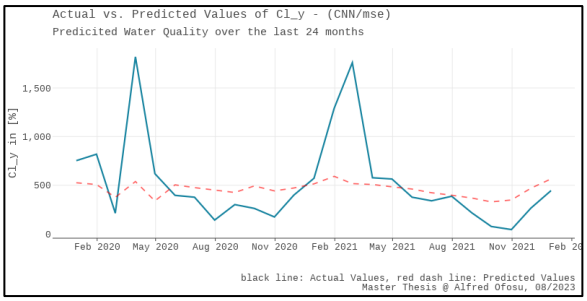


Figure 152 Predicted concentrations of Chloride – sLNN, 2

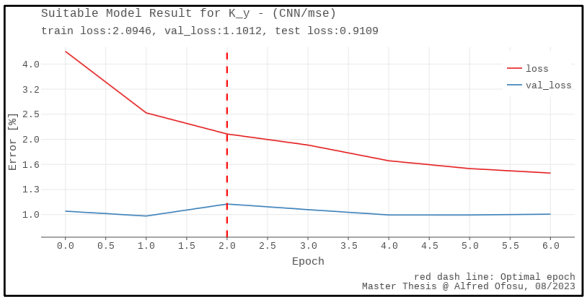


Figure 153 Modelling results, Potassium – sLNN, 2

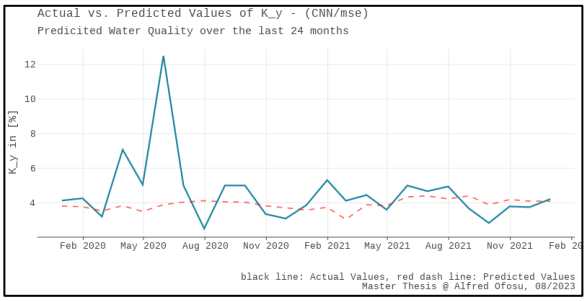


Figure 154 Predicted concentrations of Potassium – sLNN, 2

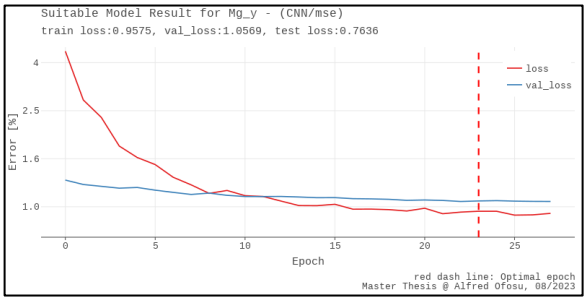


Figure 155 Modelling results, Magnesium – CNN, 2

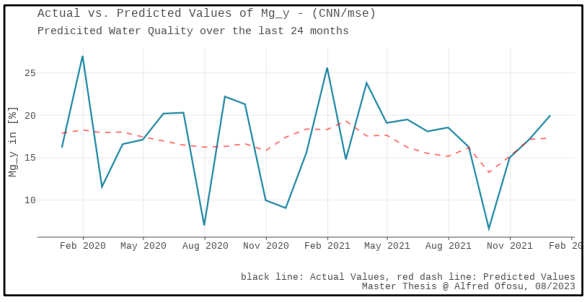


Figure 156 Predicted concentrations of Magnesium – CNN, 2

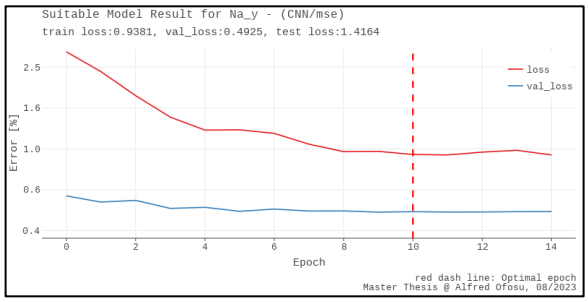


Figure 157 Modelling results, Sodium – CNN, 2

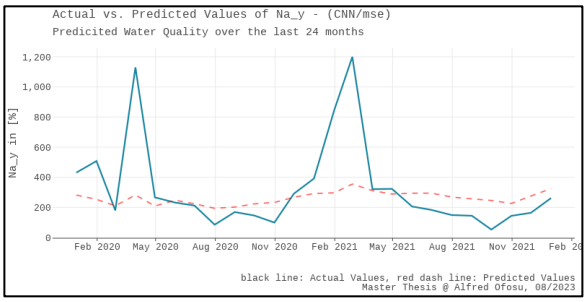


Figure 158 Predicted concentrations of Sodium – CNN, 2

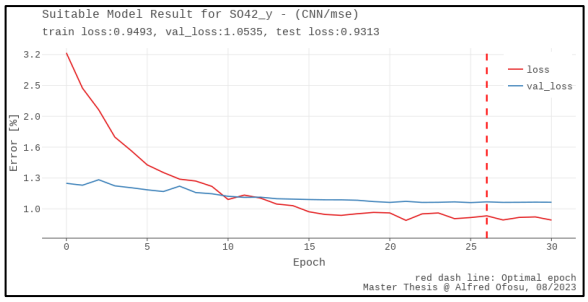


Figure 159 Modelling results, Sulphate – CNN, 2

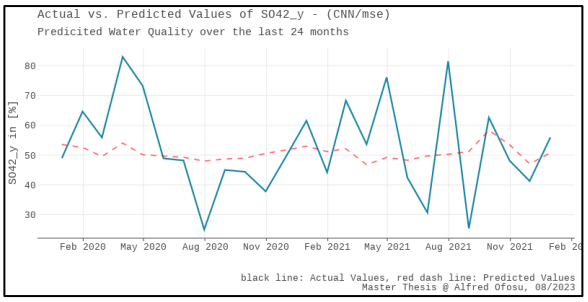


Figure 160 Predicted concentrations of Sulphate – CNN, 2

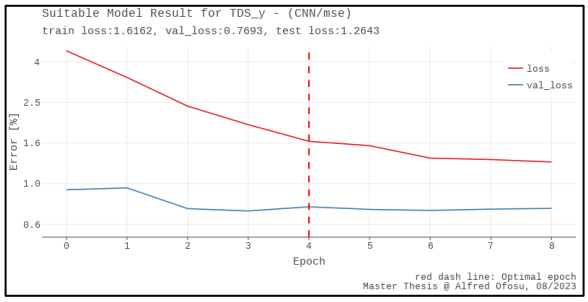


Figure 161 Modelling results, TDS – CNN, 2

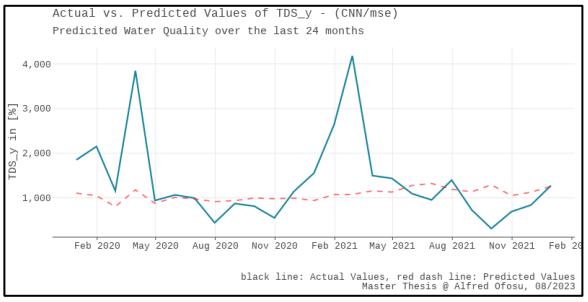


Figure 162 Predicted concentrations of TDS – CNN, 2

The results of the computations on the recurrent neural network (DNN) model are presented below. The figures on the left exhibit the training loss (indicated by the red line), validation loss (denoted by the blue line), and the point of optimal epoch (marked by the red vertical dashed line). Meanwhile, the figures on the right display the predictions made on the unobserved test dataset.

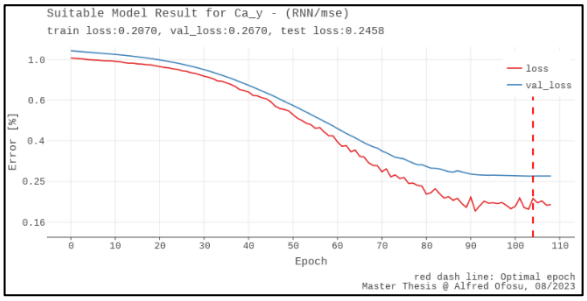


Figure 163 Modelling results, Calcium – RNN, 2

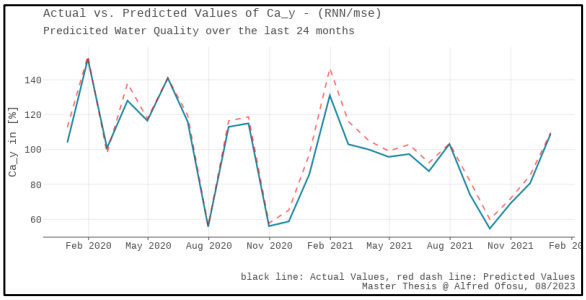


Figure 164 Predicted concentrations of Calcium – RNN, 2

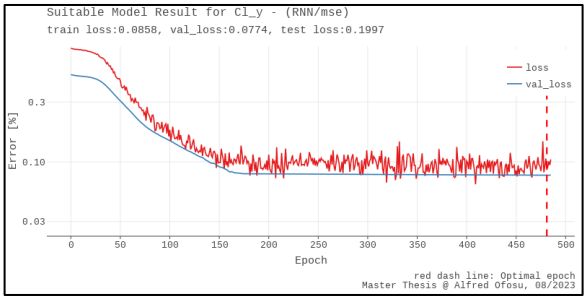


Figure 165 Modelling results, Chloride – RNN, 2

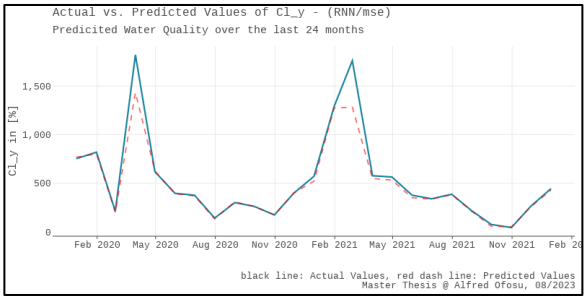


Figure 166 Predicted concentrations of Chloride – RNN, 2

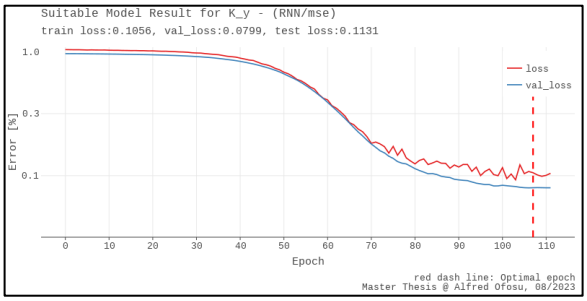


Figure 167 Modelling results, Potassium – RNN, 2

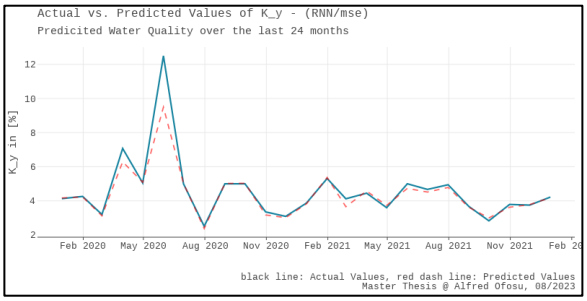


Figure 168 Predicted concentrations of Potassium – RNN, 2

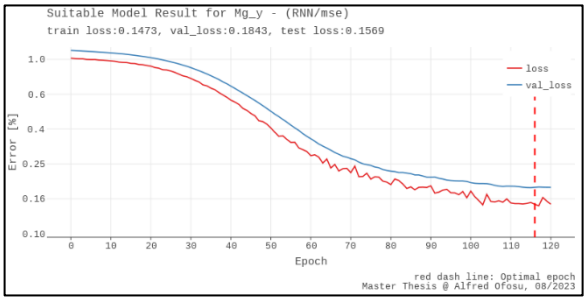


Figure 169 Modelling results, Magnesium – sLNN, 2

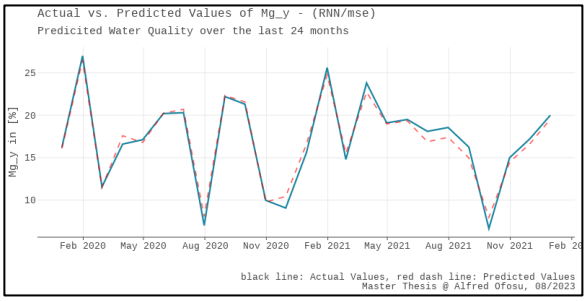


Figure 170 Predicted concentrations of Magnesium – sLNN, 2

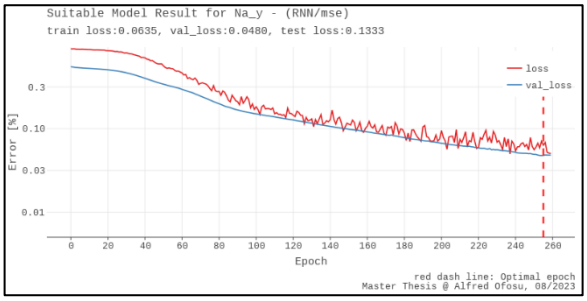


Figure 171 Modelling results, Sodium – sLNN, 2

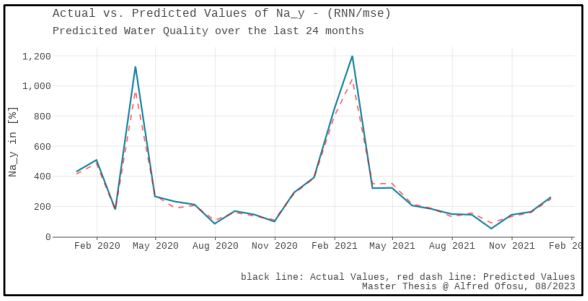


Figure 172 Predicted concentrations of Sodium – sLNN, 2

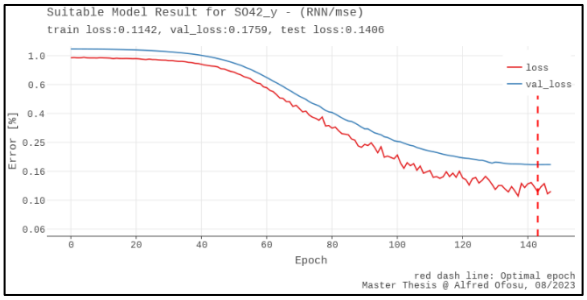


Figure 173 Modelling results, Sulphate – RNN, 2

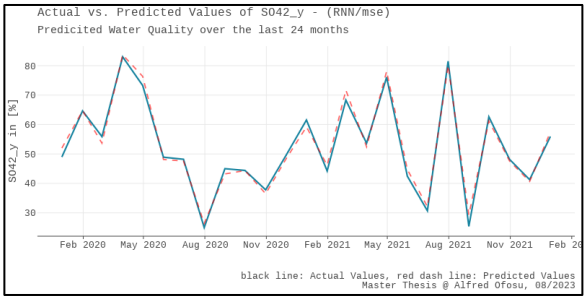


Figure 174 Predicted concentrations of Sulphate – RNN, 2

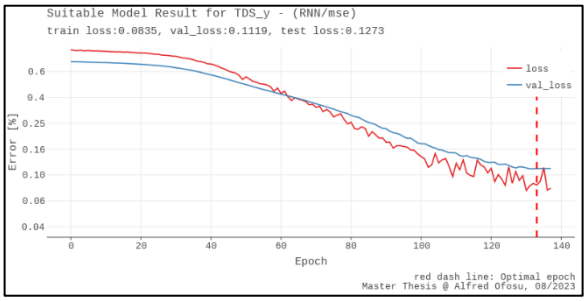


Figure 175 Modelling results, TDS – RNN, 2

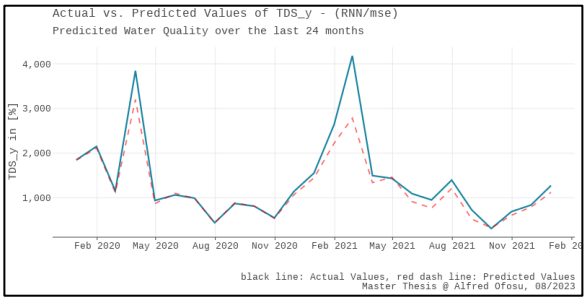


Figure 176 Predicted concentrations of TDS – RNN, 2

4.3.3 Water Quality Index Results

Imputed values were used to calculate the CCMEQWI for each month using Equation 9, Equation 10, Equation 11, Equation 12, Equation 13, and Equation 14. A DNN model was applied to predict the water quality index for the next month (t+1). The Don River watershed received indices between fair and excellent as shown in Table 4.

Table 4 Count of WQI categories received by the Don River watershed monthly.

WQI Categories	Range	Count	Percentage
Excellent	95 – 100	1288	63%
Good	80 – 94	208	10%
Fair	65 – 79	514	25%
Marginal	45 – 64	42	2%
Poor	0 – 44	0	0%

Overall, based on the outcome of these results, the Don River exhibited a favorable monthly water quality, with over 60% of the water falling into the "excellent" quality category. The water quality standards used to derive these indices were defined by specific thresholds: Calcium (≤1000 mg/L), Chloride (≤110 mg/L), pH (between 6.5 and 8.5), Sodium (≤200 mg/L), Sulphate (≤500 mg/L), Temperature (≤15°C), and Total Dissolved Solids (TDS) (≤500 mg/L).

For the prediction of the water quality index, a combination of the RNN model and a DNN, designed with an autoencoder architecture, was employed. The hyperparameters used for the model in Figure 178 were as follows:

- RNN model: 1 layer with 16 units
- DNN model: 3 layers with 32, 16, and 32 units, respectively
- Activation function: Exponential Linear Unit (ELU)
- Optimizer: Adam with a learning rate of 1e-4
- Loss function: Mean Squared Error (MSE)

This configuration allowed for effective water quality index prediction, leveraging the strengths of both the RNN and DNN models in capturing the temporal dynamics and complex patterns inherent in the water quality data. The data used in the approach were water quality parameters, hydrometric parameters, meteorological parameters, and failed water quality indicators calculated using CCMEIWQI. Additionally, the data was scaled using the Minmax scaler which scaled the values between 0, 1.

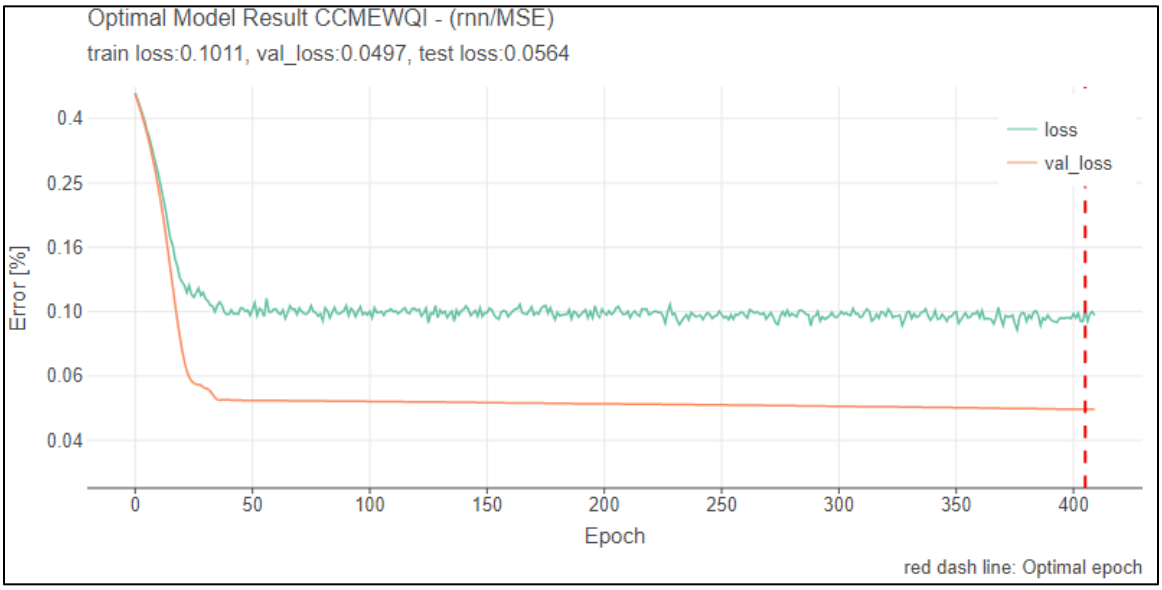


Figure 177 Model results for predicted CCME Water Quality Index

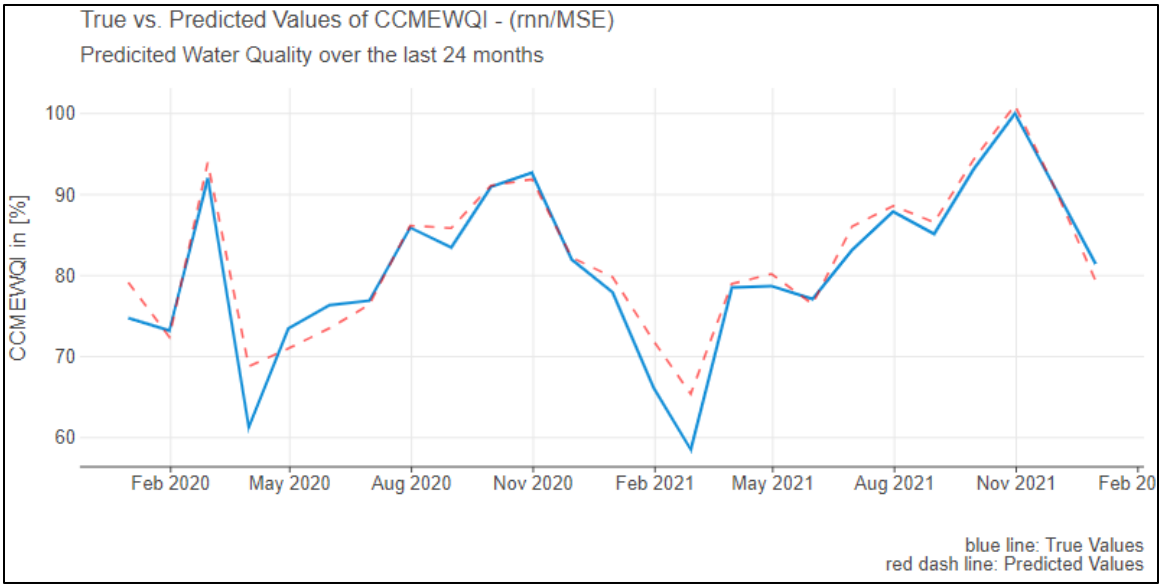


Figure 178 Predicted CCME Water Quality Index

In Figure 180, the same architecture was used, however notable changes were introduced. All hydrometric and meteorological parameters were excluded from the model. Additionally, the data was scaled using the standard scaler. The discernible disparity in results can be attributed to the differences in scaling ranges and the utilization of the mean squared error (MSE) as the loss function. MSE possesses the characteristic of amplifying the magnitude of deviations between predicted values and observed values, particularly in instances of extreme values or outliers within the dataset. With the min-max scaler effectively compressing the range of data values to a uniform scale between 0 and 1, the deviations between predicted and observed values were contained within a relatively small range. This compression essentially led to squared errors that were significantly smaller, to the point where they could approach or even reach zero. This was not case for the standard scaler, which removes the mean and scales the values to a unit variance:

$$Z = \frac{(X - \mu)}{\sigma} \qquad \textit{Equation 19}$$

where X is the sample, μ is the mean, and σ the standard deviation.

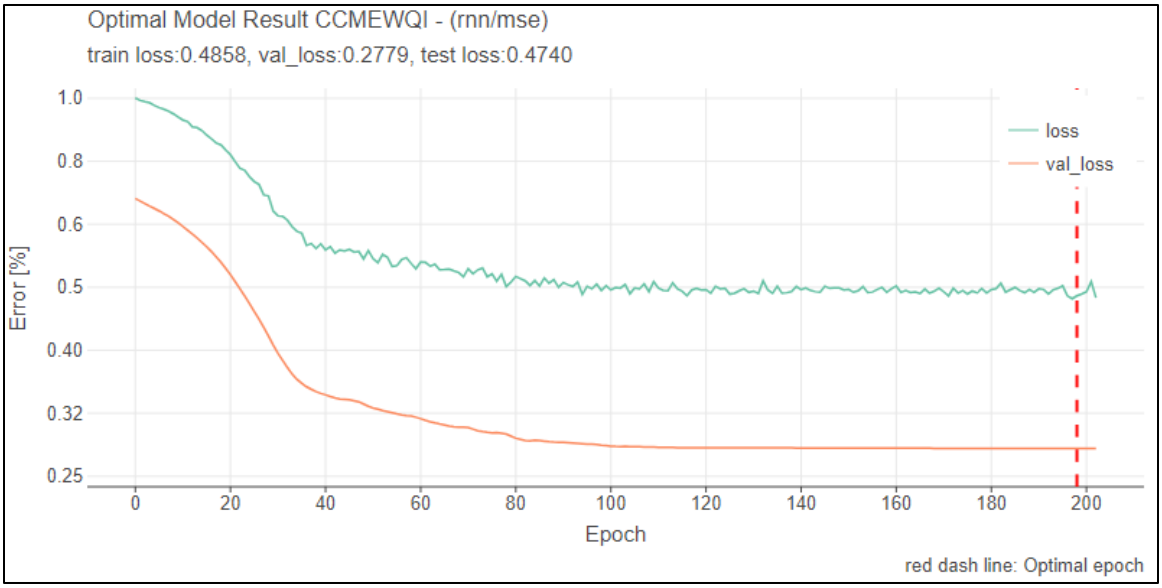


Figure 179 Model results for predicted CCME Water Quality Index - 2

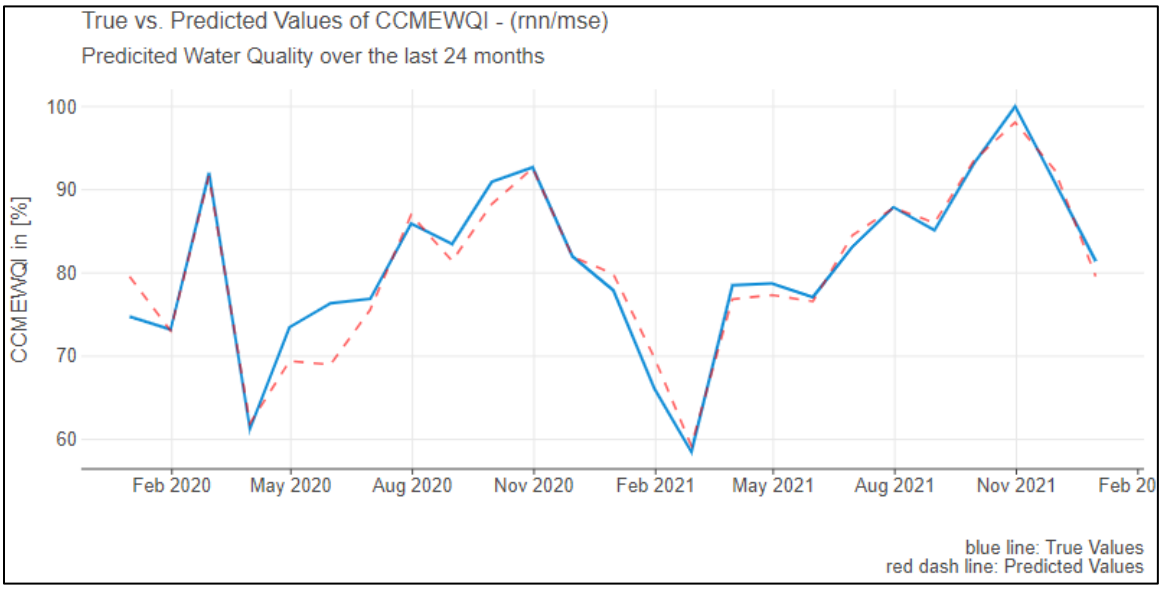


Figure 180 Predicted CCME Water Quality Index - 2

5 Conclusion

The investigation aimed to address RQ 1: "Which data-driven model provides reliable and efficient predictions of major ions in natural resources?" By extensively comparing various model architectures, the modelling results revealed that combining the Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) models was the most suitable option for predicting major ion concentrations. These models demonstrated the capability to capture intricate patterns within the environmental time series data, thereby showcasing their effectiveness in this context.

RQ 2 inquired about the impact of introducing imputed values in a sparse dataset within the data-driven water quality prediction framework. Meticulous analysis found that the choice of imputation methods had a minimal effect on the overall outcomes. All methods for imputing a few consecutive missing values provided similar results; however, when the interval of missing values increased, the iterative imputer resembled the seasonality and dynamics of the time series over the missing interval better.

RQ 3 delved into the impact of introducing exogenous parameters into the data-driven water quality prediction framework. The investigation revealed that exogenous parameters, particularly hydrological factors, were vital in predicting major ion concentrations; including these external parameters enhanced the model's prediction accuracy, especially when dealing with the complexities of natural resource data.

In conclusion, this study demonstrates the applicability of deep learning models in deriving meaningful insights from even scarce data sets. DL techniques have proven to be a robust approach for modelling water quality, particularly in scenarios where simulation models cannot be used due to the absence of detailed information on inter-water body processes and their watersheds.

It's important to emphasize that the cornerstone of successful modelling lies in the data itself. The number of observations should be substantially increased to comprehensively capture the intricate dynamics of water quality parameters within a water column. Obtaining datasets with observation frequencies reflecting seasonal and interannual variations of water quality parameters would enable the models to better capture the system's nuances and yield more accurate predictions.

It was observed that scaling methods held substantial importance. Depending on the specific problem and the need to consider outliers, a suitable scaler that does not eliminate these outliers should be chosen. In this regard, employing the Mean Squared Error (MSE) as a loss function was recommended.

This research points to the importance of feature selection for reliable predictions of water quality parameters. It suggests that correlation analysis or hydrologic knowledge does not guarantee that the selected features can improve model performance. These issues should be investigated further.

The Don River watershed is subjected to significant anthropogenic influence, which consequently impacts the concentrations of major ions. Nonetheless, comprehensive data concerning land use alterations, population growth, and specific human activities with the necessary level of detail for the timespan between 1964 and 2021 were not found leaving this important aspect of assessment uninvestigated; thus, it becomes imperative to delve into this critical aspect and conduct a thorough investigation in the future.

The models devised within this research were designed to reflect the patterns inherent in the available raw observational data. Nonetheless, the existing observation frequencies need to catch up in detecting daily or weekly fluctuations in the magnitudes of water quality parameters. To achieve this goal, higher-frequency observations are imperative. Furthermore, the models developed are site-specific. However, when applied to datasets

collected on other watersheds, the suggested framework for data analysis, preprocessing, and modelling will result in reliable models.

By addressing these aspects, DL models can be powerful tools in comprehending and predicting water quality dynamics, eventually contributing to more effective environmental management strategies.

6 Bibliography

Adu, K., Yu, Y., Cai, J., Asare, I. & Quahin, J. (2022). The influence of the activation function in a capsule network for brain tumor type classification. *International Journal of Imaging Systems and Technology*, *32*(1), 123–143. https://doi.org/10.1002/ima.22638

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. https://doi.org/10.1155/2020/6659314

Alley, E. R. (2007). *Water quality control handbook* (2nd ed., Vol. 2). McGraw-Hill. https://doi.org/10.1036/0071467602

Averill, B. A. & Eldredge, P. (2011). *Principles of General Chemistry*. Saylor Foundation. https://2012books.lardbucket.org/books/principles-of-general-chemistry-v1.0/s31-appendix-c-dissociation-consta.html

Bartos, T. T. & Ogle, K. M. (2002). Water Quality and Environmental Isotopic Analyses of Ground-Water Samples Collected from the Wasatch and Fort Union Formations in Areas of Coalbed Methane Development—Implications to Recharge and Ground-Water Flow, Eastern Powder River Basin, Wyoming. U.S. Department of the Interior, U.S. Geological Survey. https://pubs.usgs.gov/wri/wri024045/

Boccadoro, P., Daniele, V., Gennaro, P. D., Lofù, D. & Tedeschi, P. (2022). Water quality prediction on a Sigfox-compliant IoT device: The road ahead of WaterS. *Ad Hoc Networks*, *126*, 102749-.

Bordens, K. S. & Abbott, B. B. (2022). *Research design and methods: A process approach.* McGraw-Hill Education.

Brittain, J., Cendon, M., Nizzi, J. & Pleis, J. (2018). Data scientist's analysis toolbox: Comparison of Python, r, and sas performance. *SMU Data Science Review*, *1*, 7.

CCME. (2017). Water Quality Index User's Manual 2017 Update.

Cho, K., Merrienboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv*. https://doi.org/10.48550/arxiv.1406.1078

Chollet, F. (2022). *Deep Learning with Python, Second Edition*. Manning Publications. https://www.manning.com/books/deep-learning-with-python-second-edition?a_aid=keras

Clevert, D.-A., Unterthiner, T. & Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)v. *Arxiv*. https://arxiv.org/pdf/1511.07289

Colaco, B., Jackeray, S., Doradla, A. S. & Rane, R. (2020). A comparative review between programming tools used in data science. *IJCRT*.

Duhayyim, M. A., Mengash, H. A., Aljebreen, M., Nour, M., Salem, N. M., Zamani, A. S., Abdelmageed, A. A. & Eldesouki, M. I. I. (2022). Smart Water Quality Prediction Using Atom Search Optimization with Fuzzy Deep Convolutional Network. *Sustainability (Basel, Switzerland)*, *14*(24), 16465-.

Eaton, A. D., Clesceri, L. S., Rice, E. W. & Greenberg, A. E. (2005). *Standard methods for the examination of water and wastewater* (Vol. 21). Co-Published by American Public Health Association, American Water Works Association, and Water Environment Federation.

ECCC, E. and C. C. C. (2022, 14. September). *Watershed Hydrology and Ecology Research*. https://www.canada.ca/en/environment-climate-change/services/water-overview/science/watershed-hydrology-and-ecology-research.html#toc2

Hendrycks, D. & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *Arxiv*. https://arxiv.org/pdf/1606.08415.pdf

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*. https://doi.org/10.48550/arxiv.1207.0580

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Klambauer, G., Unterthiner, T., Mayr, A. & Hochreiter, S. (2017). Self-Normalizing Neural Networks. *ArXiv*. https://doi.org/10.48550/arxiv.1706.02515

Kumar, A., Tripathi, V. K., Sachan, P., Rakshit, A., Singh, R. M., Shukla, S. K., Pandey, R., Vishwakarma, A. & Panda, K. C. (2022). *Ecological Significance of River Ecosystems*. 187–202. https://doi.org/10.1016/b978-0-323-85045-2.00011-x

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. ICML.

Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press. http://neuralnetworksanddeeplearning.com/

Ofosu, A. & Erechtchoukova, M. G. (2023). Exploratory analysis of water quality in a small urbanized watershed using deep learning. In J. Vaze, C. Chilcott, L. Hutley & S. M. Cuddy (Eds.),

MODSIM2023, 25th International Congress on Modelling and Simulation (p. 653). Modelling and Simulation Society of Australia and New Zealand. https://doi.org/10.36334/modsim.2023.ofosu

Olah, C. (2015). *Understanding LSTM Networks*. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Omernik, J. M., Griffith, G. E., Irish, J. T. & Johnson, C. B. (2018). *Alkalinity and Water | U.S. Geological Survey*. https://www.usgs.gov/special-topics/water-science-school/science/alkalinity-and-water

Ozgur, C., Colliau, T., Rogers, G. & Hughes, Z. (2021). MatLab vs. Python vs. R. *Journal of Data Science*, *15*(3), 335–371. https://doi.org/10.6339/jds.201707_15(3).0001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Popkin, B. P. (1973). *Ground-water resources of Hall and eastern Briscoe Counties* (p. 85). Texas: Texas Water Development Board Report.

Saalidong, B. M., Aram, S. A., Otu, S. & Lartey, P. O. (2022). Examining the dynamics of the relationship between water pH and other water quality parameters in ground and surface water systems. *PLoS ONE*, *17*(1), e0262117. https://doi.org/10.1371/journal.pone.0262117

Shelton, J. L., Jubb, A. M., Saxe, S. W., Attanasi, E. D., Milkov, A. V., Engle, M., Freeman, P. A., Shaffer, C. A. & Blondes, M. S. (2021). Machine Learning Can Assign Geologic Basin to Produced Water Samples Using Major Ion Geochemistry. *Natural Resources Research*, *30*(6), 4147–4163. https://doi.org/10.1007/s11053-021-09949-8

Simulink, M. &. (2022). *What Is Deep Learning? | How It Works, Techniques & Applications - MATLAB & Simulink*. https://www.mathworks.com/discovery/deep-learning.html

Spellman, F. R. (2008). *Handbook of Water and Wastewater Treatment Plant Operations*. https://doi.org/10.1201/9781420075311

SSEA. (2002). *Tributary Water Quality – Severn Sound Environmental Association*. https://www.severnsound.ca/programs-projects/monitoring/tributary-water-quality/

TRCA. (2018). *Don River Watershed | TRCA Watershed and Ecosystem Reporting*. https://reportcard.trca.ca/watershed-report-cards/don-river/

TRCA & AECOM. (2017). *Don River Hydrology Update*. https://trca.ca/app/uploads/2020/01/Don_Hydrology-Study-Update-_FINAL-12-18.pdf

USGS. (2023). *Major Ions*. https://www.usgs.gov/labs/national-water-quality-laboratory/science/science-topics/major-ions

Wu, X., Zhang, Q., Wen, F. & Qi, Y. (2022). A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China. *Water (Basel)*, *14*(21), 3408-.

Yao, S., Zhang, Y., Wang, P., Xu, Z., Wang, Y. & Zhang, Y. (2022). Long-Term Water Quality Prediction Using Integrated Water Quality Indices and Advanced Deep Learning Models: A Case Study of Chaohu Lake, China, 2019–2022. *Applied Sciences*, *12*(22), 11329-.

Zhang, C., Wenna Zhang, Huang, Y. & Gao, X. (2017). Analysing the correlations of long-term seasonal water quality parameters, suspended solids and total dissolved solids in a shallow reservoir with meteorological factors. *Environmental Science & Pollution Research*, *24*(7), 6746–6756.

Zhou, S., Song, C., Zhang, J., Chang, W., Hou, W. & Yang, L. (2022). A Hybrid Prediction Framework for Water Quality with Integrated W-ARIMA-GRU and LightGBM Methods. *Water (Basel)*, *14*(9), 1322-.

7 Appendix

7.1 Useful links

1. Link to ArcGIS Map: https://arcg.is/01yDCP
2. Python code implementation: https://github.com/alfredofosu/python.projects

7.2 Abbreviations

ANN: Artificial Neural Network

CCMEI: Canadian Council of Ministers of the Environment

CNN: Convolutional Neural Network

DL: Deep Learning

DNN: Deep Neural Network

ECCC: Environment Climate Change Canada

LSTM: Long Short-Term Memory

ML: Machine Learning

MSC: Meteorological Service Canada

NN: Neural Network

PWQMN: Provincial (Stream) Water Quality Monitoring Network

RNN: Recurrent Neural Network

sLNN: Single Layer Neural Network

TRCA: Toronto Region and Conservation Authority

WSC: Water Survey of Canada

7.3 Additional Figures and Tables

Table 5 List of Stations

ID	Station	Lat	Long	Domain	Start date	End date	Status	Agency
1	6008500102	43.65	-79.35	water quality	1964-01-01	1995-12-31	I	PWQMN
2	6008500202	43.76	-79.43	water quality	1965-01-01	1993-12-31	I	PWQMN
3	6008500302	43.80	-79.40	water quality	1966-01-01	2021-12-31	A	TRCA and PWQMN
4	6008500402	43.80	-79.48	water quality	1966-01-01	2021-12-31	A	TRCA and PWQMN
5	6008500502	43.85	-79.43	water quality	1964-01-01	1988-12-31	I	PWQMN
6	6008500602	43.87	-79.43	water quality	1965-01-01	1967-12-31	I	PWQMN
7	6008500702	43.87	-79.43	water quality	1965-01-01	1966-12-31	I	PWQMN
8	6008500802	43.87	-79.43	water quality	1966-01-01	1966-12-31	I	PWQMN
9	6008500902	43.88	-79.43	water quality	1965-01-01	1967-12-31	I	PWQMN
10	6008501002	43.86	-79.43	water quality	1965-01-01	1967-12-31	I	PWQMN
11	6008501102	43.85	-79.41	water quality	1965-01-01	1967-12-31	I	PWQMN
12	6008501202	43.86	-79.43	water quality	1966-01-01	1967-12-31	I	PWQMN
13	6008501302	43.68	-79.37	water quality	1972-01-01	1991-12-31	I	PWQMN
14	6008501402	43.69	-79.36	water quality	1979-01-01	2021-12-31	A	TRCA and PWQMN
15	6008501802	43.66	-79.35	water quality	2000-01-01	2000-12-31	I	PWQMN
16	DM 6.0	43.70	-79.33	water quality	2015-01-01	2021-12-31	A	TRCA

17	DN008WM	43.80	-79.38	water quality	2015-01-01	2021-12-31	A	TRCA
18	HY003	43.67	-79.42	Precipitation	2005-04-29	2020-11-10	I	TRCA
19	HY008	43.69	-79.37	Precipitation	2009-06-23	2021-03-10	A	TRCA
20	HY016	43.68	-79.32	Precipitation	2006-06-23	2020-12-09	I	TRCA
21	HY017	43.79	-79.47	Discharge & Stage	2012-12-05	2017-07-10	I	TRCA
22	HY018	43.74	-79.39	Discharge & Stage	2013-09-25	2021-03-10	A	TRCA
23	HY019	43.69	-79.36	Stage	2007-07-10	2021-03-10	A	TRCA
24	HY021	43.83	-79.48	Precipitation	2005-04-27	2021-03-10	A	TRCA
25	HY022	43.76	-79.35	Stage	2015-01-01	2021-03-10	A	TRCA
26	HY027	43.77	-79.46	Precipitation & Stage	2007-09-13	2021-03-10	A	TRCA
27	HY036	43.82	-79.31	Precipitation	2006-05-08	2021-03-10	A	TRCA
28	HY038	43.90	-79.61	Precipitation	2013-04-15	2020-12-04	I	TRCA
29	HY039	43.84	-79.59	Precipitation	2008-06-23	2021-03-02	A	TRCA
30	HY055	43.83	-79.60	Precipitation	2013-02-26	2020-12-11	A	TRCA
31	HY062	43.70	-79.33	Discharge & Stage	2012-11-22	2021-02-05	A	TRCA
32	HY064	43.77	-79.51	Precipitation	2013-04-15	2014-12-03	I	TRCA
33	HY068	43.73	-79.36	Stage	2012-11-22	2021-02-10	A	TRCA
34	HY069	43.92	-79.48	Precipitation	2013-04-23	2020-12-07	I	TRCA

35	HY070	43.88	-79.38	Precipitation	2013-04-23	2020-12-07	I	TRCA
36	HY079	43.66	-79.36	Stage	2011-03-14	2021-03-10	A	TRCA
37	HY080	43.73	-79.27	Stage	2015-01-01	2018-01-25	I	TRCA
38	HY085	43.72	-79.41	Precipitation	2014-10-24	2014-12-05	I	TRCA
39	HY087	43.69	-79.52	Precipitation	2015-04-15	2020-12-09	I	TRCA
40	HY092	43.80	-79.38	Discharge & Stage	2015-05-11	2021-03-01	A	TRCA
41	HY093	43.80	-79.38	Discharge & Stage	2015-05-11	2018-05-18	I	TRCA
42	HY094	43.75	-79.32	Precipitation	2015-05-29	2021-03-10	A	TRCA
43	HY100	43.79	-79.47	Discharge & Stage	2017-07-27	2021-03-10	A	TRCA
44	HY112	43.80	-79.40	Discharge & Stage	2020-04-22	2021-03-01	A	TRCA
45	HY123	43.89	-79.44	Stage	2020-07-15	2021-03-10	A	TRCA
46	02HC004	43.77	-79.37	Discharge	1965-01-01	1965-12-31	I	WSC
47	02HC005	43.74	-79.40	Discharge & Stage	1945-01-01	2023-12-31	A	WSC
48	02HC024	43.69	-79.36	Discharge	1962-01-01	2023-12-31	A	WSC
49	02HC029	43.76	-79.35	Discharge	1964-01-01	1996-12-31	I	WSC
50	02HC056	43.83	-79.44	Discharge & Stage	2005-01-01	2023-12-31	A	WSC
51	6152953	43.87	-79.38	weather	1974-07-01	1994-05-31	I	MSC

52	6154131	43.93	-79.52	weather	1979-07-01	1981-10-31	I	MSC
53	6154135	43.90	-79.62	weather	1959-03-01	1978-04-30	I	MSC
54	6154950	43.87	-79.48	weather	1962-01-01	1990-01-31	I	MSC
55	6154951	43.85	-79.52	weather	1961-07-01	1968-08-31	I	MSC
56	6154994	43.83	-79.35	weather	1961-10-01	1979-02-28	I	MSC
57	6155680	43.92	-79.60	weather	2005-10-01	2007-05-18	I	MSC
58	6157012	43.88	-79.45	weather	1959-06-01	2014-02-28	I	MSC
59	6157014	43.87	-79.43	weather	1960-05-01	1981-10-31	I	MSC
60	6157015	43.90	-79.40	weather	1989-06-01	1991-10-31	I	MSC
61	6158255	43.80	-79.42	weather	1965-07-01	2007-06-30	I	MSC
62	6158355	43.67	-79.40	weather	2002-06-04	2023-07-05	A	MSC
63	6158384	43.75	-79.27	weather	1962-05-01	1975-04-30	I	MSC
64	6158385	43.72	-79.32	weather	1973-06-01	1984-05-31	I	MSC
65	6158386	43.73	-79.50	weather	1957-11-01	1981-04-30	I	MSC
66	6158398	43.77	-79.52	weather	1960-03-01	1979-08-31	I	MSC
67	6158406	43.65	-79.35	weather	1980-04-01	1993-12-31	I	MSC
68	6158408	43.78	-79.32	weather	1967-09-01	1980-08-31	I	MSC
69	6158409	43.86	-79.37	weather	2018-10-29	2023-07-05	A	MSC
70	6158420	43.78	-79.33	weather	1963-10-01	1966-05-31	I	MSC
71	6158422	43.75	-79.28	weather	1957-11-01	1965-04-30	I	MSC

72	6158443	43.75	-79.48	weather	1956-10-01	1982-06-30	I	MSC
73	6158474	43.72	-79.48	weather	1951-01-01	1970-11-30	I	MSC
74	6158480	43.80	-79.35	weather	1965-11-01	1975-10-31	I	MSC
75	6158536	43.68	-79.27	weather	1957-02-01	1976-11-30	I	MSC
76	6158550	43.75	-79.42	weather	1958-01-01	1969-04-30	I	MSC
77	6158567	43.70	-79.45	weather	1953-01-01	1966-02-28	I	MSC
78	6158575	43.67	-79.32	weather	1966-04-01	1981-07-31	I	MSC
79	6158578	43.62	-79.35	weather	1994-12-29	1998-03-07	I	MSC
80	6158718	43.77	-79.48	weather	1973-06-01	1987-04-30	I	MSC
81	6158731	43.68	-79.63	weather	2013-06-13	2023-07-05	A	MSC
82	6158732	43.72	-79.35	weather	1985-05-01	1987-04-30	I	MSC
83	6158733	43.68	-79.63	weather	1937-11-01	2013-06-13	I	MSC
84	6158734	43.80	-79.40	weather	1968-05-01	1969-10-31	I	MSC
85	6158740	43.80	-79.55	weather	1965-05-01	1988-06-30	I	MSC
86	6158751	43.70	-79.34	weather	2007-05-24	2015-07-31	I	MSC
87	6158752	43.80	-79.42	weather	1971-11-01	1975-03-31	I	MSC
88	6158762	43.68	-79.45	weather	1957-10-01	1981-03-31	I	MSC
89	6158765	43.72	-79.23	weather	1958-09-01	1965-12-31	I	MSC
90	6158779	43.72	-79.38	weather	1962-08-01	1993-02-28	I	MSC
91	6158830	43.77	-79.42	weather	1953-11-01	1970-03-31	I	MSC

92	6158846	43.73	-79.43	weather	1953-07-01	1973-08-31	I	MSC
93	6159510	43.95	-79.43	weather	1960-07-01	1979-03-31	I	MSC
94	61583FL	43.70	-79.48	weather	1980-11-01	1987-04-30	I	MSC
95	61584J2	43.72	-79.35	weather	1981-08-01	1988-11-30	I	MSC
96	61587PG	43.78	-79.35	weather	1973-06-01	1987-04-30	I	MSC
97	61587PP	43.65	-79.37	weather	1966-04-01	1979-11-30	I	MSC
98	61587PR	43.75	-79.25	weather	1965-07-01	1965-09-30	I	MSC
99	6158M1K	43.78	-79.32	weather	1971-04-01	1980-06-30	I	MSC
100	615HHDF	43.75	-79.38	weather	1973-06-01	1987-04-30	I	MSC
101	615HMAK	43.86	-79.37	weather	1986-05-01	2015-05-20	I	MSC
102	615S001	43.78	-79.47	weather	1994-11-01	2022-06-10	A	MSC

ALKT

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	785	Minimum	50.9
Distinct (%)	49.7%	Maximum	529.5
Missing	1057	Zeros	0
Missing (%)	40.1%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	208.42643	Memory size	41.2 KiB

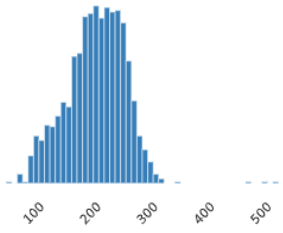


Figure 181 Statistical Analysis of Alkalinity (raw data)

ALKT

Real number (ℝ)

HIGH CORRELATION

Distinct	1784	Minimum	62.843636
Distinct (%)	86.9%	Maximum	448.48105
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	207.5858	Memory size	8.1 KiB

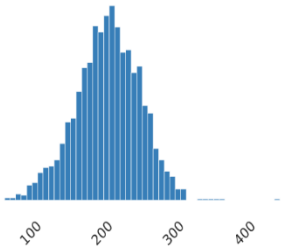


Figure 182 Statistical Analysis of Alkalinity (imputed data)

Ca2+

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	360	Minimum	0.113
Distinct (%)	48.2%	Maximum	214
Missing	1888	Zeros	0
Missing (%)	71.7%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	101.68677	Memory size	41.2 KiB

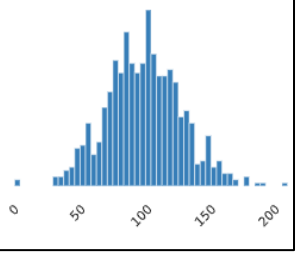


Figure 183 Statistical Analysis of Calcium (raw data)

Ca2+

Real number (ℝ)

HIGH CORRELATION

Distinct	1911	Minimum	0.113
Distinct (%)	93.1%	Maximum	194.76315
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	102.60707	Memory size	8.1 KiB

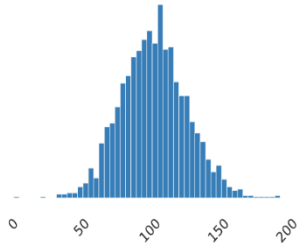


Figure 184 Statistical Analysis of Calcium (imputed data)

Cl-

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	896	Minimum	2
Distinct (%)	39.2%	Maximum	4570
Missing	347	Zeros	0
Missing (%)	13.2%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	292.466	Memory size	41.2 KiB

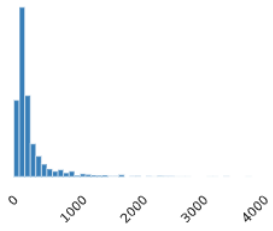


Figure 185 Statistical Analysis of Chloride (raw data)

Cl-

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	1659	Minimum	0
Distinct (%)	80.8%	Maximum	3920
Missing	0	Zeros	58
Missing (%)	0.0%	Zeros (%)	2.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	356.71716	Memory size	8.1 KiB

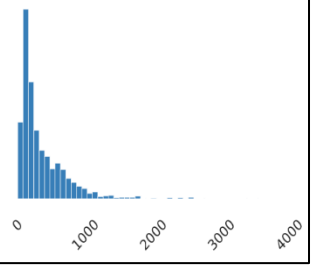


Figure 186 Statistical Analysis of Chloride (imputed data)

HARD

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	318	Minimum	1
Distinct (%)	41.0%	Maximum	665
Missing	1859	Zeros	0
Missing (%)	70.6%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	325.9058	Memory size	41.2 KiB

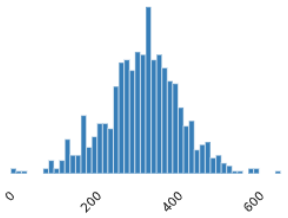


Figure 187 Statistical Analysis of Hardness (raw data)

HARD

Real number (ℝ)

HIGH CORRELATION

Distinct	1914	Minimum	1
Distinct (%)	93.3%	Maximum	621.63275
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	329.95163	Memory size	8.1 KiB

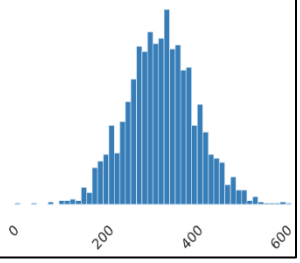


Figure 188 Statistical Analysis of Hardness (imputed data)

Mg2+

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	305	Minimum	0.0201
Distinct (%)	40.9%	Maximum	33.7
Missing	1889	Zeros	0
Missing (%)	71.7%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	17.144534	Memory size	41.2 KiB

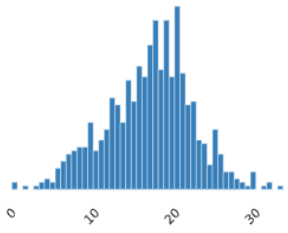


Figure 189 Statistical Analysis of Magnesium (raw data)

Mg2+

Real number (ℝ)

HIGH CORRELATION

Distinct	1903	Minimum	0.020099999
Distinct (%)	92.7%	Maximum	32.3811
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	17.31008	Memory size	8.1 KiB

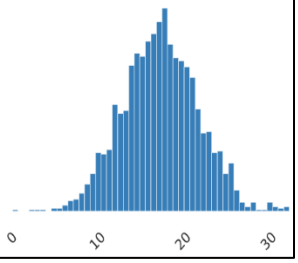


Figure 190 Statistical Analysis of Magnesium (imputed data)

Na+

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	358	Minimum	0.0051
Distinct (%)	62.6%	Maximum	2960
Missing	2063	Zeros	0
Missing (%)	78.3%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	328.14802	Memory size	41.2 KiB

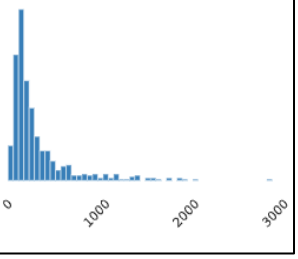


Figure 191 Statistical Analysis of Sodium (raw data)

Na+

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	1941	Minimum	0
Distinct (%)	94.6%	Maximum	2090
Missing	0	Zeros	55
Missing (%)	0.0%	Zeros (%)	2.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	208.8503	Memory size	8.1 KiB

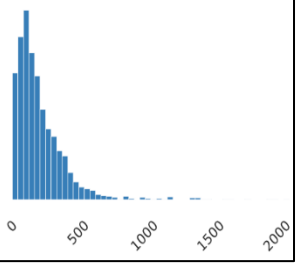


Figure 192 Statistical Analysis of Sodium (imputed data)

SO42-

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	353	**Minimum**	6.61
Distinct (%)	70.5%	**Maximum**	478.41
Missing	2134	**Zeros**	0
Missing (%)	81.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	52.253673	**Memory size**	41.2 KiB

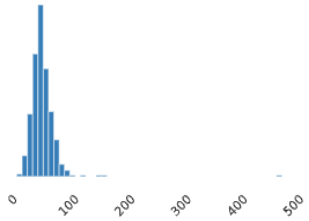


Figure 193 Statistical Analysis of Sulphate (raw data)

SO42-

Real number (ℝ)

Distinct	2020	**Minimum**	5.9142642
Distinct (%)	98.4%	**Maximum**	178.30333
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	49.451203	**Memory size**	8.1 KiB

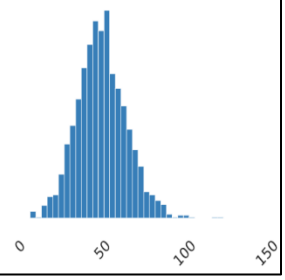


Figure 194 Statistical Analysis of Sulphate (imputed data)

Temp

Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	327	**Minimum**	-1.2
Distinct (%)	13.7%	**Maximum**	27
Missing	254	**Zeros**	33
Missing (%)	9.6%	**Zeros (%)**	1.3%
Infinite	0	**Negative**	15
Infinite (%)	0.0%	**Negative (%)**	0.6%
Mean	10.402351	**Memory size**	41.2 KiB

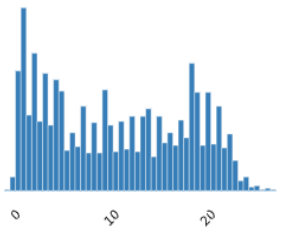


Figure 195 Statistical Analysis of Water Temperature (raw data)

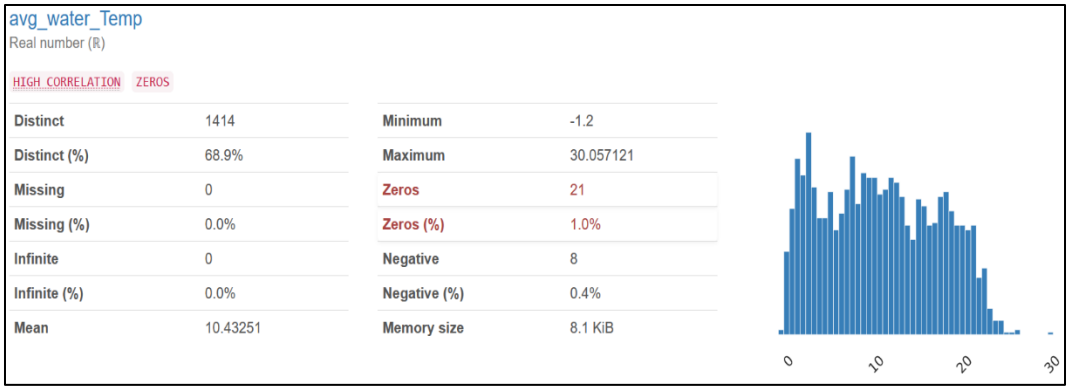


Figure 196 Statistical Analysis of Water Temperature (imputed data)

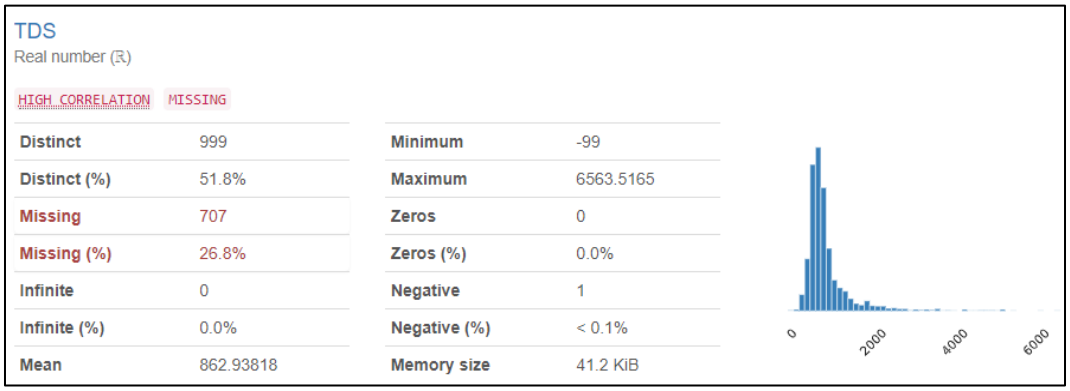


Figure 197 Statistical Analysis of Total Dissolved Solids (raw data)

The minimum value -99 in Figure 197 is an erroneous value.

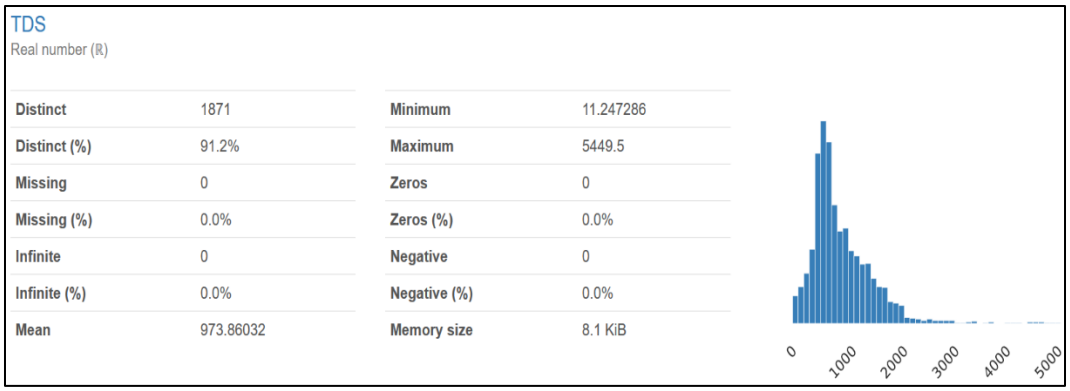


Figure 198 Statistical Analysis of Total Dissolved Solids (imputed data)

Table 6 Hyper parameter optimization results -– Baseline model, Research Approach 1

Major Ion	Units	Activation function	Dropout	Val loss	Learning Rate
Calcium	512	selu	No	0.0966	0.00002
Chloride	512	relu	Yes	0.0572	0.00958
Potassium	512	relu	Yes	0.0592	0.00773
Magnesium	32	relu	Yes	0.1054	0.00096
Sodium	512	gelu	Yes	0.0613	0.00008
Sulphate	1024	gelu	No	0.0756	0.00002
TDS	32	gelu	Yes	0.0803	0.00026

Table 7 Hyper parameter optimization results – Baseline model, Research Approach 2

Major Ion	Units	Activation function	Dropout	Val loss	Learning Rate
Calcium	32	elu	0.2	0.1069	0.08123
Chloride	96	relu	0.5	0.0.507	0.00668
Potassium	32	elu	0.5	0.10013	0.08123
Magnesium	96	elu	0.5	0.10038	0.02662
Sodium	128	elu	0.5	0.06139	0.00375
Sulphate	32	elu	0.2	0.07939	0.08123
TDS	32	relu	0.2	0.08292	0.01032

Table 8 Hyper parameter optimization results - DNN

Major Ion	Layers	Units	Activation functions	Dropout	Val loss	Learning Rate
Calcium	4	128, 256, 32	tanh, elu, selu, relu	No	0.0899	0.00002
Chloride	2	64, 64	gelu, relu	No	0.0553	0.01144
Potassium	4	256, 32, 1024, 1024	gelu	Yes	0.0590	0.00124
Magnesium	2	64, 512	gelu, relu	No	0.1109	0.00034
Sodium	3	64, 128, 1024	selu, gelu, relu	No	0.0626	0.00002
Sulphate	5	18, 64, 256, 32, 32	gelu, elu, elu, gelu, elu	No	0.0741	0.00169
TDS	2	32, 32	tanh, selu	Yes	0.0797	0.00843

Table 9 Hyper parameter optimization results - CNN

Major Ion	Layers	Units	Activation functions	Dropout	Val loss	Learning Rate
Calcium	3	1024, 64, 32	elu, tanh, gelu	Yes	0.0965	0.00008
Chloride	3	64, 256, 1024	relu,elu, relu	No	0.0554	0.00979
Potassium	4	256, 256, 32, 256	tanh, tanh, relu, tanh	Yes	0.0586	0.02263
Magnesium	5	32, 32, 512, 64, 1024	Selu, tanh, selu, tanh, gelu	No	0.1048	0.00689
Sodium	2	32, 32	tanh, selu	Yes	0.0636	0.00843
Sulphate	3	512, 128, 1024	selu, gelu, gelu	Yes	0.0763	0.00002
TDS	4	512, 128, 64, 256	gelu, selu, selu, gelu	Yes	0.0800	0.00093

Table 10 Hyper parameter optimization results - RNN

Major Ion	Layers	Units	Dropout	Val loss	Learning Rate
Calcium	3	256, 1024, 128	No	0.0925	0.00047
Chloride	4	32, 512, 128, 256	No	0.0557	0.00012
Potassium	1	64	Yes	0.0577	0.00084
Magnesium	1	64	Yes	0.10565	0.00029
Sodium	1	512	Yes	0.06105	0.00017
Sulphate	1	32	Yes	0.07364	0.0015
TDS	1	32	No	0.07988	0.00751

Table 11 Statistical analysis before and after imputation of missing values.

Parameter	Ca_2^+		Cl^-		K^+		Na^+	Mg_2^+		SO_4^{2-}		TDS	
Statistics	B[9]	A[10]	B	A	B	A	B	A	B	A	B	A	B
Min	0.1	0.1	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.6	6.0	49.0
5-th %	62.1	64.4	56.0	40.1	1.7	1.2	72.6	16.7	8.7	10.1	28.2	26.0	380.0
Q1	86.0	86.0	113.5	124.8	2.6	2.6	129.8	85.1	14.4	14.4	41.2	39.0	542.0
Median	106.0	102.0	164.9	221.0	3.3	3.5	203.0	156.2	17.9	17.3	49.2	49.0	688.0
Q3	124.0	117.8	291.0	488.9	4.2	4.5	375.3	273.8	21.0	20.2	60.0	59.0	922.0
95-th %	154.0	142.3	960.7	1005.8	6.0	6.0	1096.5	534.1	25.9	24.6	79.8	74.0	2112.0

[9] "B" stands for **Before** the imputation of missing values.

[10] "A" stands for **After** the imputation of missing values.

Max	214.0	194.8	4570.0	3920.0	20.0	20.0	2960.0	2090.0	33.7	32.4	178.3	178.0	6564.0
Range	213.9	194.7	4566.0	3920.0	19.9	19.9	2960.0	2090.0	33.7	32.4	171.7	172.0	6515.0
IQR	38.0	31.6	177.5	364.0	1.6	2.0	245.5	188.7	6.6	5.8	18.8	19.0	380.0
Std. Dev.	28.7	24.1	392.2	374.0	1.9	1.6	357.5	205.3	5.3	4.4	17.6	15.0	679.0
Kurtosis	0.4	0.3	25.7	14.6	26.6	11.5	10.3	18.1	0.2	0.2	8.8	2.8	16.1
Mean	105.9	102.6	292.1	356.7	3.6	3.6	331.7	208.9	17.6	17.3	51.6	49.5	881.8
Skewness	0.1	0.1	4.3	3.0	3.8	1.5	2.8	3.3	-0.2	0.1	1.7	0.6	3.5
Variance	825.0	580.4	153847.0	139912.0	3.4	2.5	127776.0	42132.0	27.7	19.5	309.5	231.0	460477.0