HARMFUL ALGAL BLOOM EVENTS PREDICTION: COMPARING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

A DATA-DRIVEN APPROACH IN THE FIELD OF WATER QUALITY PREDICTION

RONG LIAO

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2088124

COMMITTEE

Dr. Bruno Nicenboim
MSc. Büşra Özgöde Yigin

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 23, 2023

WORD COUNT

8679

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my family in China and the Netherlands for their unwavering support and belief in my abilities.
I would also like to extend my heartfelt thanks to my thesis supervisor, Dr. Bruno Nicenboim, for his expertise and constructive feedback throughout this research endeavor. His contributions have played a pivotal role in shaping the direction and quality of this thesis.
Furthermore, I would like to express my gratitude to my second reader, MSc. Büşra Özgöde Yigin, for her meticulous evaluation of my thesis.

HARMFUL ALGAL BLOOM EVENTS PREDICTION: COMPARING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

A DATA-DRIVEN APPROACH IN THE FIELD OF WATER QUALITY PREDICTION

RONG LIAO

Abstract

Harmful algal bloom events (HABs), one of the water quality issues, are increasingly frequent, threatening both aquatic life and human health. Machine learning is gaining prominence in predicting and analyzing aquatic environmental data, offering new possibilities for addressing the issue of HABs. Previous studies have utilized various models to make HABs predictions, such as random forest (RF), support vector machine (SVM), and recurrent neural network (RNN), but the problem of sparsity in the dataset was seldom addressed. Most researchers dealt with missing data using imputations in the past. The distinguishing aspect of this paper's approach is the investigation of potential differences in machine learning and deep learning algorithms, not only in terms of different imputation methods but also by incorporating datasets with factual data. This study used RF as the baseline model, with support vector regression (SVR) and artificial neural network (ANN) as competing models. The models were trained using one dataset without imputation and compared with datasets that underwent three kinds of imputation techniques: median imputation, KNN imputation, and multiple imputation. The dataset contained water quality parameters measured in New York Harbor over a century, with a shape of 89021 by 100. We found that it was feasible to increase HABs prediction efficiency with factual data and domain knowledge. The SVR model with median imputation performed the best, with an R^2 score of 0.859, a root mean squared error (RMSE) score of 9.360, and a mean absolute error (MAE) score of 5.737 after feature selection. Furthermore, machine learning and deep learning algorithms using data with and without imputations generated similar outputs, but the deep learning algorithm incurred higher computational costs. Finally, 17 most important features were identified for contributing to the occurrence of HABs in New York

Harbor, including fluorometer, pH, and dissolved oxygen (DO). Notably, several novel predictors, such as fluorometers, nitrate/nitrite, and ammonium, have emerged as promising variables for HABs prediction in our study. However, the imputation methods utilized in our research were not sufficient when the features were entirely absent.

1 SOURCE/CODE/ETHICS/TECHNOLOGY STATEMENT

The dataset was acquired from the publicly accessible New York City Open Data portal. The work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership before, during, and after the completion of this study. All figures in this thesis belong to the author. Some parts of the code have been adapted from a publicly available source, which can be found at https://www.freecodecamp.org/news/machine-learning-pipeline. The reused or adapted code fragments are clearly indicated in the notebook. For convenient access to the aforementioned notebooks, please visit https://github.com/bunnybunny1120/master_thesis_2023. The author received language assistance in refining the paper's content using a generative language model (ChatGPT-3.5), which included paraphrasing, spell checking, and grammar correction. No other typesetting tools or services were employed.

2 INTRODUCTION

The notion of water quality is complex and influenced by various factors. There are primarily four factors that can be used to determine water quality. Firstly, physical parameters such as water turbidity, total suspended solids (TSS), and electrical conductivity (EC) can impact water quality. For example, cloudy water is considered unsuitable for drinking, and the presence of certain metals in water determines its suitability for irrigation or firefighting. Secondly, chemical factors such as pH, DO, and chemical oxygen demand (COD) can affect the survival of aquatic organisms. High levels of acidity or alkalinity in water may indicate chemical or industrial pollution. Thirdly, anthropogenic factors resulting from human activities such as irrigation, extraction, and household waste have a significant impact on the water system. Lastly, biological factors such as the presence of bacteria, algae, and viruses can also contribute to changes in water quality. Algae, for instance, can not only cause odor and taste issues in water but also create more significant problems by producing toxins (Akhtar et al., 2021; Cariappa, 2004; Kisi et al., 2023; Ortenberg & Telsch, 2003; Patrick, 1973; Summers, 2020). Although HABs are natural occurrences that have been documented throughout history, non-toxic blooms can also cause anoxic conditions and harm aquatic life and the most concerning issue for humans is the consumption of seafood contaminated by algal species that produce neurotoxins. This can result in gastrointestinal and neurological illnesses (Hallegraeff, 2010; Nick et al., 2019).

When predicting HABs, the specific parameter Chlorophyll-a (Chl-a) is used as the target. Chlorophylls are a class of pigments found in photosynthetic organisms such as plants and algae. They enable these organisms to convert sunlight into chemical energy through a process called photosynthesis. Among the three types of chlorophyll (a, b, and c) found in phytoplankton, chlorophyll-a is typically used as a measure of phytoplankton biomass and production, which contributes to the development of HABs (Blankenship, 2014; Glibert et al., 2018; Jeffrey & Vesk, 1997). Consequently, the occurrence of HABs can serve as an indicator of water pollution, as they are triggered by conditions associated with the presence of Chlorophyll-a.

3 research goal & problem statement

The goal of this research is to determine whether factual data can yield comparable outcomes to imputed data when predicting HABs and if machine learning and deep learning algorithms perform differently in this context. If substantiated, the utilization of data without imputation has the potential to enhance the efficiency of HABs prediction by saving time and mitigating potential biases arising from the imputation process. Moreover, given the increasing occurrence of HABs, which poses a critical global concern, the development of such efficient methodologies becomes imperative for addressing this pressing issue within a specified time frame (Anderson et al., 2021; Masó & Garcés, 2006).

Thanks to the rapid increase in data volume in recent years, machine learning has gained significant importance in the analysis, classification, and prediction of time-series data related to the aquatic environment (M. Zhu et al., 2022). However, the issue of high sparsity in water quality datasets has received relatively little attention (Ma et al., 2020). This issue will be addressed in Section 4. Based on current understanding, there is a lack of previous studies that have investigated the potential discrepancies between datasets without imputations and imputed datasets when using machine learning methods for water quality predictions. Consequently, the understanding of how machine learning and deep learning models perform differently in such settings is yet to be explored.

3.1 *Societal & Scientific Relevance*

The societal relevance of this study lies in its ability to facilitate early detection of pollution events, enabling governments to prioritize resources in water quality management and prevent potential ecological disasters. Moreover, the utilization of machine learning-based water quality analysis aids in safeguarding industries such as fisheries, aquaculture, and tourism, which face various challenges, including the outbreak of diseases linked to poor water quality (Burford, 1997; Burkholder, 1998; Cruz et al., 2021; L.-H. Lee & Lee, 2015; Ojea et al., 2023). From a scientific perspective, applying machine learning techniques can reveal previously undetectable patterns, leading to a deeper understanding of complex aquatic ecosystems. Furthermore, future researchers aiming to enhance their models for water environment studies can use the features identified in this research as a reference, and the method employed in this paper has the potential to expedite the setup of their experiments.

3.2 *Research Questions*

Based on the context established from the previous sections, the main research question of this thesis can be formulated as follows:

> *How to predict harmful algal bloom events in New York Harbor by using machine learning and deep learning algorithms with and without data imputations?*

The following supporting questions could be formulated in order to respond to the primary study question:

- SQ1 To what extent do machine learning and deep learning models perform differently with factual data and data undergone imputation with multiple techniques?
- SQ2 Which are the most important water quality indicators that contribute to harmful algal bloom events in New York Harbor?

3.3 *Findings*

The study compared different models with various imputation techniques for predicting HABs using data from a specific water system. The results of the study indicated the following: 1) The utilization of domain knowledge enhanced the efficiency of HABs prediction. 2) SVR with median imputation achieved the highest R^2 score of 0.859, the RMSE score of 9.360, and the

MAE score of 5.737 after feature selection. 3) The performance of models using factual data did not significantly deviate from those using imputation techniques. 4) Machine learning and deep learning algorithms yielded comparable results, albeit with the deep learning approach exhibiting a longer computational runtime. 5) Seventeen important parameters, including fluorometers, pH, and temperature, were identified for predicting HABs in New York harbor. Notably, certain features such as fluorometers, nitrate/nitrite, and ammonium appeared to be underutilized in previous HABs prediction studies. However, 6) when the features were entirely absent, they tended to have no predictive power with imputation methods used in this research.

4 RELATED WORK

In the field of water quality prediction, a substantial body of literature has investigated the application of machine learning techniques. This section is divided into four parts. Firstly, machine learning in water quality prediction is introduced. Secondly, innovative methods in HABs are discussed. Thirdly, examples of utilizing small datasets in water quality prediction are provided. The last part reveals the practice of using domain knowledge and the data-driven nature of this field. Finally, the research gap is identified.

4.1 *Machine Learning used in Water Quality Prediction*

Water quality prediction is the process of estimating the levels or concentrations of various parameters, such as DO, biochemical oxygen demand (BOD), and total phosphorus (TP) in a water body.

In order to lower cost and find more efficient ways to make predictions, some studies utilized regression models to predict the water quality index (WQI). Asadollah et al. (2021) proposed extra tree regression (ETR) as opposed to SVR and decision tree regression (DTR) using only factual data, but the alternative methods such as imputation techniques were not explored. Chen et al. (2020) evaluated the performance of ten machine learning models such as logistic regression (LR), linear SVM, and RF using big data (33,612 samples), and they found that it was beneficial to use large datasets and that ensemble models outperformed traditional models. Furthermore, in some studies focusing on regression tasks within the realm of machine learning, RF emerged as the most efficacious model (Castrillo & García, 2020; Koranga et al., 2022).

However, traditional machine learning methods might have limitations in their ability to capture the non-linear or changing dynamics in various water bodies. A few comparative studies have shown that ANNs have demonstrated better efficiency and accuracy compared to SVM, simple multiple linear regression (MLR), SVR, and RF (Azrour et al., 2022; Djerioui et al., 2019; Ooi et al., 2021; S. Zhu et al., 2019). Conversely, in certain studies, the superiority of deep learning algorithms, specifically ANNs, did not manifest when compared to SVM or RF (Haghiabi et al., 2018; Koranga et al., 2022).

While single shallow models may fall short in capturing the significant relationships and patterns that exist over extended periods, more sophisticated models have been employed in the prediction of water quality parameters. These include the integration of the wavelet function in ANN and long short-term memory (LSTM) models, as well as the combination of recurrent neural networks (RNN) with Dempster-Shafer theory (RNNs-DS), and the fusion of grey theory with ANNs such as backpropagation neural network (BPNN), radial basis function neural network (RBFNN), and generalized regression neural network (GRNN) (L. Li et al., 2019; Zamani et al., 2023; Zhai et al., 2019). Although these studies have predominantly focused on exploring advanced models to improve predictive capabilities, it is important to note that they did not explicitly address the underlying issue of a dataset characterized by a substantial proportion of missing values.

4.2 *Predicting HABs State of Art*

More related to this study, various innovative techniques have been developed for the prediction of HABs. These include the merged-LSTM model proposed by Cho and Park (2019), which integrated multiple data sources and improved the performance of ANN and two-layer LSTM models. Derot et al. (2020) introduced a coupling model that utilized a long-term database spanning 34 years, employing K-means clustering for unsupervised learning and RF with a sliding window for supervised learning. Image-based approaches have also been explored, such as the algal morphology deep neural network (AMDNN) by Yuan et al. (2023), enabling real-time processing and differentiation of algae species on-site. Additionally, Guo et al. (2021) combined the underwater Imaging FlowCytobot (IFCB) with RF for real-time classification of HABs, achieving performance comparable to the convolutional neural network (CNN) with transfer learning. Despite the utilization of state-of-the-art models in HABs prediction, the major drawback of such sophisticated methods might be their potential inefficiency under limited time or resource constraints. Moreover, these studies did not address a prevalent issue encountered in datasets with a high percentage of missing values, especially in large datasets spanning several decades.

4.3 *Predicting HABs using Small Datasets*

The primary objective of this study is to investigate the effectiveness of utilizing factual data compared to data with imputations in predicting HABs. However, it is important to acknowledge that removing missing values from the dataset may result in a reduced amount of available data for analysis. Nevertheless, existing research provides compelling evidence supporting the feasibility of using machine learning methods for HABs prediction, regardless of the dataset size. For example, Yu et al. (2021) demonstrated the utility of AdaBoost, ANN, GBDT, and SVM on two relatively small datasets—one with 365 samples and the other with 40 samples—from distinct locations in the United States and China. Similarly, Shin et al. (2020) employed SVR, Bagging, RF, XGBoost, and LSTM with a two-year dataset from the Nakdong River in South Korea. However, it should be noted that Yajima and Derot (2017) used different datasets from lakes and reservoirs in Japan and applied RF with a sliding window strategy to predict HABs. They found that the limited data volume could potentially constrain the predictive power of the model.

4.4 *Domain Knowledge and Data Driven Nature*

In the context of evaluating regression models, previous research has widely employed evaluation metrics such as R^2, RMSE, MAE. Notably, among machine learning and deep learning models, RF, SVM, and ANN have frequently been utilized in comparative studies (Azrour et al., 2022; Deng et al., 2021; Djerioui et al., 2019; Haghiabi et al., 2018; Koranga et al., 2022; S. Lee & Lee, 2018; L. Li et al., 2019; Ly et al., 2021; Ooi et al., 2021; Zamani et al., 2023; Zheng et al., 2021). Based on their effectiveness demonstrated in previous literature, their relevance to HABs prediction, and their suitability for comparative analysis, we chose RF, SVR, and ANN as competing models and R^2, RMSE, and MAE to be the evaluation metrics.

Furthermore, the feature selection step played a crucial role in HABs research due to the influence of hydrological and geographical variations. For instance, Ly et al. (2021) employed time series models, regression models, deep learning models, and adaptive neuro-fuzzy inference system (ANFIS), using a dataset from the Han River in South Korea spanning a period of 10 years. They concluded that meteorological factors such as precipitation, current flow rate, and temperature significantly impact the prevalence of HABs in the region due to the monsoon-like climate.

Additionally, insights drawn from prior investigations could provide a foundation for establishing initial models. For example, Deng et al. (2021) set up SVM and ANN models with a dataset spanning a 30-year time period in Hong Kong Tolo Harbour, based on several features derived from previous research in the same water system. They identified that BOD, TIN, DO, phosphate (PO4), and pH were the key factors inducing the occurrence of HABs. Consistently, about six features were commonly mentioned having relations with Chl-a in previous studies: nitrogen (Cho & Park, 2019; Jeong et al., 2022; K.-M. Kim & Ahn, 2022; Z. Li et al., 2023; Ly et al., 2021; Xia et al., 2020; Yajima & Derot, 2017), BOD (Deng et al., 2021; S. Lee & Lee, 2018; Ly et al., 2021; Shin et al., 2020), pH (Cho & Park, 2019; K.-M. Kim & Ahn, 2022; S. Lee & Lee, 2018; Ly et al., 2021; Wang & Xu, 2020; Yajima & Derot, 2017), phosphorus (Cho & Park, 2019; K.-M. Kim & Ahn, 2022; Z. Li et al., 2023; Ly et al., 2021; Yajima & Derot, 2017), temperature (Cho & Park, 2019; Jeong et al., 2022; K.-M. Kim & Ahn, 2022; S. Lee & Lee, 2018; Ly et al., 2021; Shin et al., 2020; Wang & Xu, 2020) and DO (Cho & Park, 2019; Deng et al., 2021; K.-M. Kim & Ahn, 2022; S. Lee & Lee, 2018; Shin et al., 2020; Wang & Xu, 2020). Importantly, a pertinent case study conducted in the New York Harbor, which is identical to our research setting's water system, was dedicated to the prediction of BOD5 (Ma et al., 2020). In this study, Chl-a was integrated as one of the predictors, thereby suggesting a plausible association between Chl-a and BOD5.

However, the establishment of a standardized set of predictors that can effectively predict various water bodies poses a challenge, as each water system exhibits unique characteristics. Additionally, discrepancies in the collection of water quality parameters by relevant entities such as the local government or regulatory bodies, may also contribute to this challenge. For instance, certain features identified in previous studies for Chl-a prediction such as COD (Cho & Park, 2019; S. Lee & Lee, 2018; Shin et al., 2020; Yajima & Derot, 2017), Cyanobacteria (S. Lee and Lee, 2018) and solar radiation (Cho and Park, 2019) are absent in our dataset. Consequently, the available quantity and types of data play a crucial role in determining the effectiveness of HABs predictions.

4.5 *Research Gap*

The current work builds upon and improves previous lines of inquiry by addressing a crucial issue that has been overlooked in most studies: the high percentage of missing values in the dataset. While prior investigations have frequently employed deletion or imputation methods, such as univariate imputation, KNN imputation, and multiple imputation, to handle missing values (Asadollah et al., 2021; Chou et al., 2018; Kang & Park, 2021; H.-R. Kim et al., 2022; S. Lee & Lee, 2018; Ly et al., 2021; Ooi et al., 2021; Shamsuddin et al., 2022; Shin et al., 2020; Yajima & Derot, 2017; Yu et al., 2021), there is a lack of knowledge in the comparative investigation regarding the utilization of factual data versus imputed data with machine learning and deep learning algorithms in HABs prediction. The motivation of addressing this knowledge gap is to contribute to the field of HABs prediction in New York Harbor by improving the efficiency of predictions, especially in scenarios where missing values are prevalent, and resources and time are limited. To address this objective, we will conduct a comparative analysis of several ML models using both imputed and non-imputed datasets, select the best model to identify most important predictors and discuss the implications of these findings in the subsequent section 8.

5 METHODOLOGY

This chapter outlines the methodologies employed to yield the essential outcomes for addressing the research questions at hand. To address the first sub-question, SQ1, which investigates the performance disparities between machine learning and deep learning models when using factual data versus data that has undergone imputation with multiple techniques, two distinct datasets are prepared. Furthermore, a range of models are employed, and evaluation metrics are established to compare and evaluate their performance. This process enables the identification of the most effective model for subsequent feature selection to answer the second sub-question, SQ2.

5.1 *Datasets*

In order to compare model performances with and without imputations, we prepared two datasets after data cleaning. One small dataset with only factual data (denoted as 'No Imputation' method) comprising features of domain knowledge, and another dataset for different imputations was created through utilizing the same rows as the factual dataset containing all the features. In this regard, the models with and without imputations could be compared.

5.2 *Models*

Three models were adopted as illustrated in Chapter 4.4. The RF model served as the benchmark model. SVR and ANN were the competing models.

5.2.1 *Random Forest*

RF regression was chosen as the baseline model due to its ensemble nature, which involves building multiple decision trees and aggregating their predictions through voting or averaging. The advantage of RF lies in its interpretability, attributed to the measure of feature importance, as well as its robustness to noise and outliers (Oshiro et al., 2012; Schonlau & Zou, 2020).

5.2.2 *Support Vector Regression*

SVR is an extension of SVM, and both aim to find a hyperplane that separates data points with the maximum margin. SVR specializes in predicting a continuous variable by minimizing the difference between the predicted and actual values of the output variable while maintaining a maximum allowable deviation (epsilon) from the optimal hyperplane (Drucker et al., 1997).

5.2.3 *Artificial Neural Network*

ANNs are widely used for water quality prediction. They are a type of feedforward network in supervised learning, particularly a multi-layer perceptron (MLP). MLP consists of an input layer, one or multiple hidden layer(s), and an output layer. Various activation functions could be employed, including rectified linear unit (ReLU), hyperbolic tangent (tanh), sigmoid, and softmax to introduce non-linearity and improve the model's accuracy (Sarker, 2021). In our study, we utilized the MLP with a single hidden layer.

5.3 *Imputation Techniques*

5.3.1 *Categorical Data Imputation*

Two common methods for imputing categorical data are replacing missing values with the most frequent category and creating a new category specifically for missing values. Given a substantial proportion of missing categorical data, we opted for the latter approach to mitigate the risk of favoring an erroneous category and introducing bias.

5.3.2 *Median Imputation*

We chose the median over the mean as a measure of central tendency because the median provides a more robust estimation in the presence of outliers. Additionally, since the data in our study primarily consists of continuous variables, the mode, which is suitable for categorical variables, is not preferred. The application of the median as a summary statistic is supported by previous studies in the field (Han et al., 2012; Navarro, 2018).

5.3.3 *KNN Imputation*

KNN Imputation is a machine learning algorithm that finds the K data points in the dataset most similar to the target and imputes missing values by averaging the values of the K nearest neighbors for continuous data (Batista and Monard, 2002).

5.3.4 *Multiple Imputation (MICE)*

MICE (Multivariate Imputation by Chained Equations), a multiple imputation method, involves several steps to impute missing values. It begins with a simple imputation for all columns, such as using the mean, and sets the imputed mean as missing for a specific column. The missing value is then imputed through regression, based on the other columns. This process is repeated for all columns. Multiple cycles are conducted with updated imputations until the difference in imputed values between iterations is minimized (Azur et al., 2011).

5.4 *Evaluation Metrics*

As discussed in Chapter 4.4, three evaluation metrics were chosen for this study: R^2, RMSE and MAE, based on which the best model will be chosen using validated results.

5.4.1 *Coefficient Determination (R^2)*

The first evaluation metric is coefficient determination (R^2). It ranges from 0 to 1, quantifying the degree to which variations in the output can be explained by changes in the independent variables. The higher value indicates better model performance (James et al., 2021, p. 70).

5.4.2 *Root Mean Square Error (RMSE)*

The second evaluation metric is the root mean square error (RMSE). It has been frequently used in evaluating numerical models in the fields of meteorology, air quality, and climate research. One of its advantages lies in its ability to maintain the same units as the original data, thereby facilitating meaningful interpretation. However, in some cases, RMSE may be susceptible to outliers (Hodson, 2022).

5.4.3 *Mean Absolute Error (MAE)*

To mitigate the sensitivity of outliers, the mean absolute error (MAE) can be employed as a third evaluation metric. The MAE is particularly suitable when errors follow a Laplacian distribution. Previous research (Hodson, 2022) suggests that when variables have a normal distribution, the resulting errors are also likely to exhibit normality.

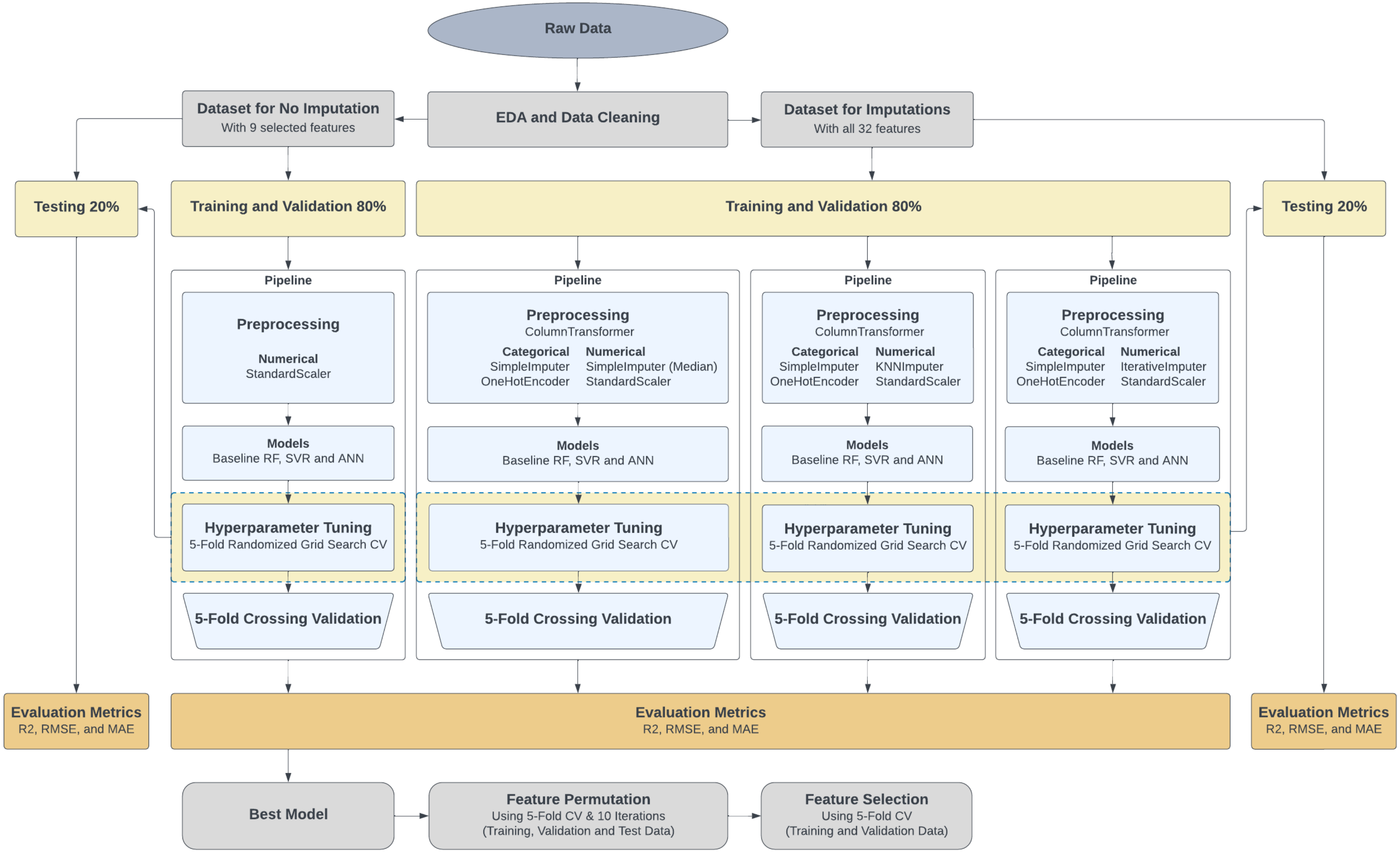
5.5 *Feature Permutation*

By addressing the second sub-question, SQ2, the focus shifts towards identifying the key features that are vital for predicting HABs in New York Harbor. The model that exhibits the highest performance, as determined in SQ1, is utilized to conduct feature importance permutation, enabling the establishment of a ranking system for the predictors. This facilitates the selection of the most significant features associated with HABs prediction.

6 EXPERIMENTAL SETUP

This section provides a comprehensive workflow for the project, depicted in Figure 1. The workflow initiates with a description of the raw dataset, followed by exploratory data analysis (EDA), data cleaning and preprocessing. Subsequently, the detailed process of experimental procedure is outlined. Lastly, the robustness of the methodology is discussed, along with the actual implementation process.

Figure 1: Flowchart Steps from Raw Data to Feature Selection



6.1 *Raw Dataset Description*

The dataset used in this study originates from New York City (NYC) Open Data. It is collected through the Harbor Survey Program, which has been monitoring water quality in New York Harbor since its inception in 1909. This dataset is widely utilized by regulators, scientists, educators, and citizens to assess the impact, developments, and improvements in water quality. Most features in the dataset are numeric and represent in-situ measurements of various water quality parameters. The dataset consists of 89,201 rows and 100 columns, covering the period from 1909 to 2022. The rationale for selecting this dataset lies in its reliability, substantial size, and relatively unexplored nature.

6.2 *EDA and Data Cleaning*

6.2.1 *Descriptive Statistics and Data Preparation*

The data collection in New York Harbor involved 5,087 distinct locations. Certain features included measurements from both the top water surface and the bottom ocean bed. For example, the oxygen reduction potential (ORP) had similar sample sizes for the top (13,078 samples) and bottom (12,792 samples) measurements, with comparable mean, standard deviation, minimum, and maximum values. However, some features exhibited substantial disparities in sample sizes between the top and bottom measurements. For instance, Nitrate/Nitrite had 41,472 top samples but only 1,006 bottom samples. Notably, most features demonstrated similar distribution patterns across both the top and bottom samples, as evidenced by the 25th, 50th, and 75th percentiles.

To ensure data usability, several data preparation steps were implemented. Initially, column names were modified to address length, spaces, and special characters. The columns were then manually re-indexed based on their physical, chemical, and biological attributes, as well as their data types (numerical or categorical). The target value was shifted to the last column for convenience. Redundant or invalid columns were identified and removed, including those with constant values, limited or no data entries, location information, and date information. Rows lacking a target value and columns with fewer than 1000 rows were dropped from the dataset, ensuring a minimum threshold of one thousand rows. Columns containing both top and bottom sample data were combined by calculating the average, following established practices in previous studies (Asadollah et al., 2021; Deng et al., 2021). In cases where there was a significant imbalance between the top and bottom entries of a feature, only the majority entry was retained to preserve more data.

6.2.2 *Categorical Feature Treatment*

Three columns in the dataset were categorized as categorical features: weather, wind direction, and current direction. The weather feature had two distinct states, namely "dry" and "wet." To ensure consistency, these states were transformed into a standardized format represented by 'D' and 'W', respectively. Entries containing symbols or irrelevant characters were considered as missing values.

The descriptive statistics revealed that the current direction feature had 119 unique counts, which was unexpected considering that there should only be 16 cardinal directions, such as North-northeast (NNE), Northeast (NE), and East-northeast (ENE), among others. To address the extraneous information, values that did not correspond to any of the 16 cardinal directions were treated as missing values. The same approach was applied to the wind direction feature.

6.2.3 *Outlier Analysis*

No outliers were removed from the dataset using the standard deviation method. The majority of features exhibited non-normal distributions, as evident from the histograms and boxplots (Figure 6 and Figure 5 in the Appendix), except for pH and O2. Given that the pH values appeared within a reasonable range and represented valid measurements, they were retained. Conversely, a few negative values for the percentage of O2 saturation were manually removed, as the percentage should be positive. Further discussion regarding this issue will be presented in Chapter 8.

6.2.4 *Feature Correlations*

Heatmaps were utilized to identify highly correlated features. Variables with a correlation exceeding 85% were removed to mitigate multicollinearity and enhance model performance. The impact of feature removal can be observed in Figure 7 and Figure 8 (Appendix). It is important to note that the heatmaps exclusively presented numerical values, resulting in the exclusion of three columns (weather, wind direction, and current direction). The numbers within the heatmap indicate the strength and direction of the relationships between variables. Empty spaces or blanks indicate the absence of calculated correlations for specific variable pairs. A total of eight highly correlated features were eliminated.

To explore the relationships among the three categorical features, a contingency table was employed and visualized using a grouped bar chart shown in Figure 10 (Appendix). It was observed that the current direction exhibited a consistent pattern across different weather conditions, while south, southwest, and south-southwest wind directions were more prevalent during dry weather conditions.

6.2.5 *Missing Value Analysis*

The original dataset displayed high sparsity, with more than half of the features containing up to 80% missing values in the raw dataset prior to cleaning (Figure 9 in the Appendix). After data cleaning, the dataset was reduced to 40,919 rows and 33 columns (Figure 11 in the Appendix). Among these columns, 8 had less than 20% missing values, while 10 had more than 50% missing values.

A heatmap analysis (Figure 2 in the Appendix) was conducted to examine the distribution of missing values across the dataset. It was observed that no single row was entirely free of missing values. Certain features exhibited clustered patterns of missing values, indicating shared characteristics among corresponding rows. On the other hand, some missing values were sporadically scattered throughout different rows, suggesting random occurrences within the dataset.

To compare models using factual and imputed data, a dataset consisting solely of factual data was prepared by filtering out rows with missing values. After filtering, the dataset contained 1572 rows and 9 features. Figure 3 illustrates that 13 columns had missing values exceeding 50%, with 8 of them having 0 entries. The treatment of missing data will be discussed in Chapter 6.3.

Figure 2: Heatmap of Missing Values in the Entire Dataset after Cleaning: The x-axis represents features, and the y-axis represents row indexes. The dark blue shade indicates missing values. The figure shows that the missingness does not overlap among the rows, indicating that each row has its unique set of missing values. Additionally, some features exhibit clustered patterns of missing values, suggesting shared characteristics among corresponding rows. In contrast, other missing values appear sporadically across different rows, implying randomness in their occurrence throughout the dataset.

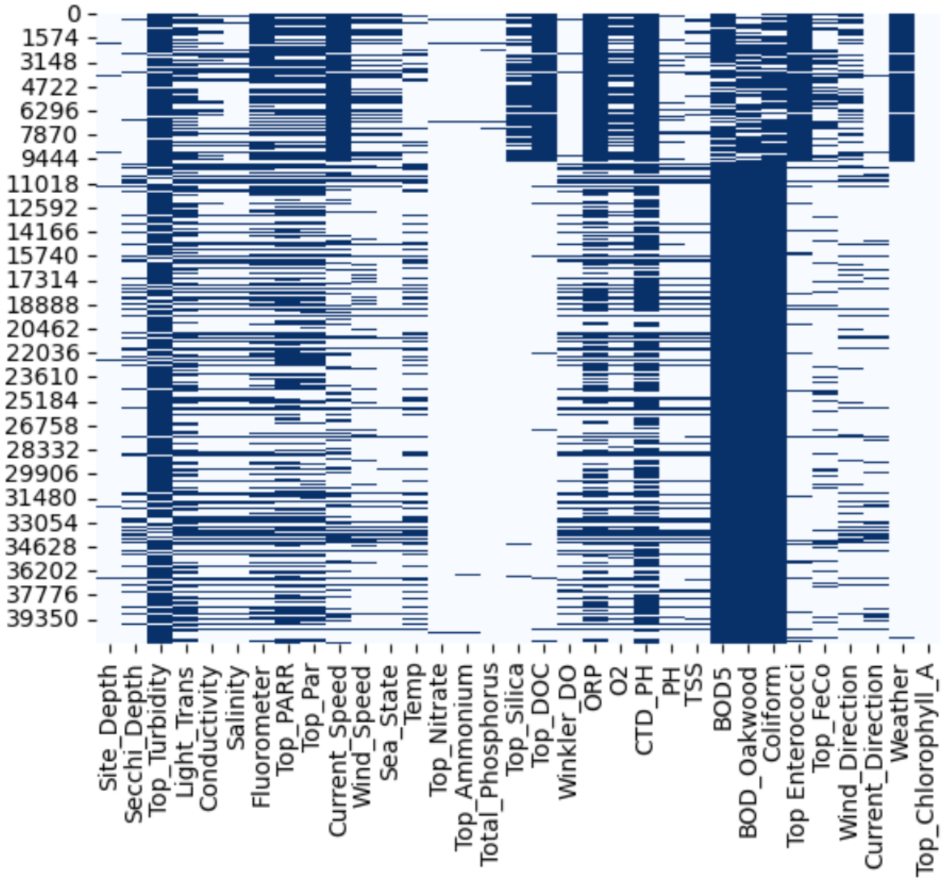
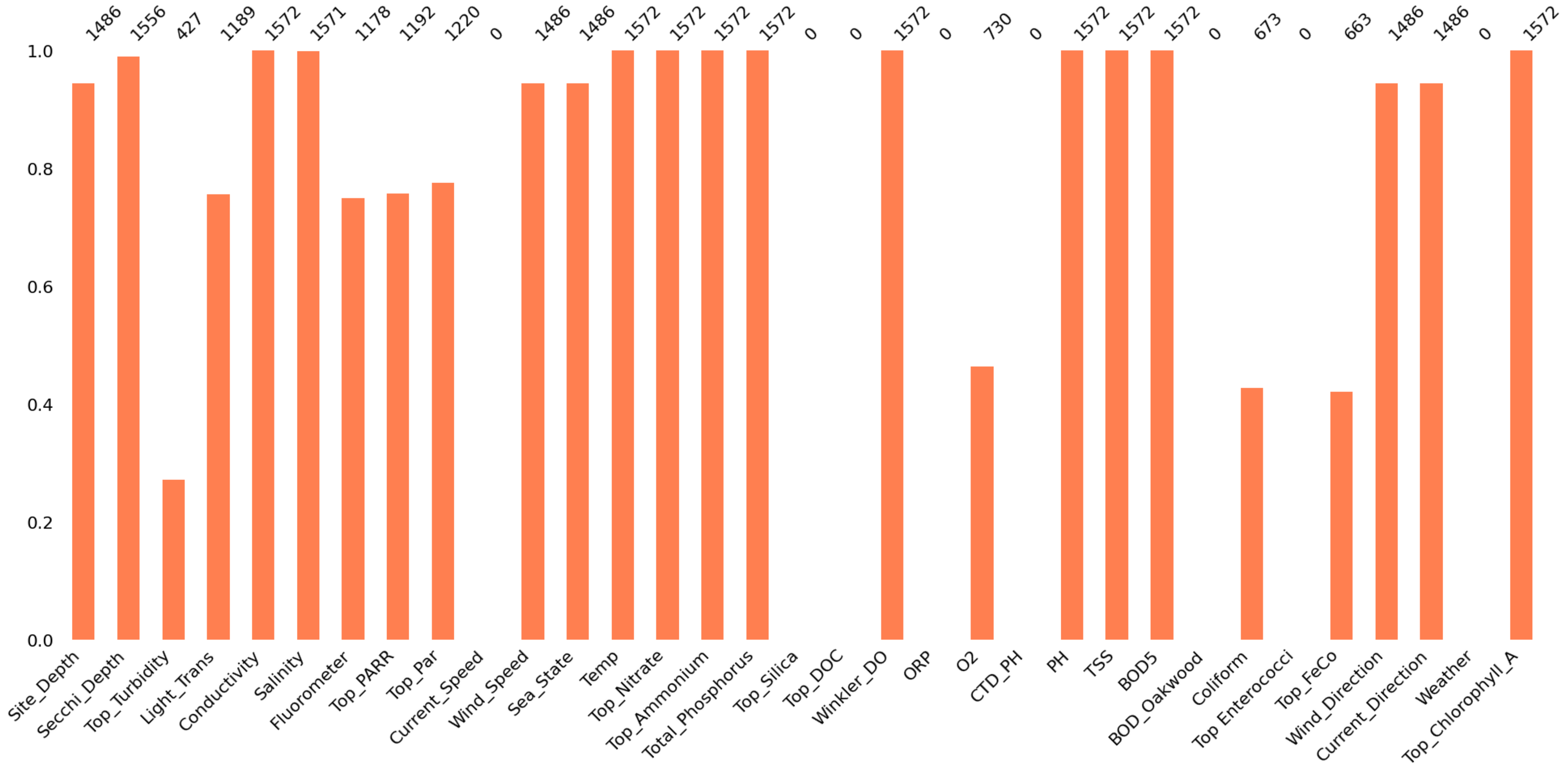


Figure 3: Missing values after cleaning. There are 1572 rows and 33 columns, including the target Chl-a. The number on top of each bar represents the actual count of the feature. Features with less than 1572 entries are imputed. The number on the left vertical axis indicates the percentage scale from 0 to 1. The numbers displayed on the top of the bars represent the count of non-null values for each corresponding feature.



6.3 *Data Preprocessing*

6.3.1 *Imputations*

We used the SimpleImputer from the scikit-learn library for median imputation of numeric data, as well as for imputing categorical data. For categorical data, instead of using the mode method, we treated missing values as a distinct category by replacing them with a constant value 'missing'. This approach allows the model to potentially capture patterns or associations related to the absence of data, without introducing bias that could arise from using the mode method.

For KNN imputation, we utilized the KNNImputer from the scikit-learn library. After experimentation, it was determined that utilizing 9 neighbors produced optimal outcomes, consistent with the observed cluster pattern of missing values discussed in Chapter 6.2.5. We used uniform weights for imputation, assuming no difference in the importance of neighbors given the manually designed column order.

For multiple imputation, we employed the IterativeImputer from the FancyImpute library. To ensure convergence, we increased the default iteration value from 10 to 50 and set the random state to 12 for reproducibility.

6.3.2 *Scaling and One-hot Encoding*

The numerical values were standardized by transforming them to have a mean of zero and a standard deviation of one. This step was important for ensuring comparability, particularly for models like SVR and ANN, which are sensitive to the scale of the data. Standardization was preferred over normalization to avoid the potential impact of outliers and to maintain the original relationships between data points.

For categorical variables, we opted for one-hot encoding instead of label encoding, because it avoids introducing unintended ordinal relationships between categories, as recommended by the scikit-learn documentation (Scikit-learn contributors, 2022).

6.4 *Pipeline*

The preprocessing steps for both numeric and categorical data were combined into a pipeline. This pipeline included a column transformer that sequentially applied the necessary preprocessing techniques to different feature subsets. Specifically, numeric features were first imputed and then scaled, while categorical features were imputed and one-hot encoded. This design ensured that each step was applied appropriately to the corresponding feature subset, maintaining consistency in our machine learning workflow.

6.5 *Experimental Procedure*

The initial stage involved data cleaning and preparation of preprocessing steps. It is worth noting that the actual preprocessing occurred after data splitting, as explained in the pipeline structure.

After data cleaning, two datasets were created: one without imputations, consisting of the factual data, and another with imputation techniques, consisting of all 32 features. The dataset containing only factual data was first created based on domain knowledge. Six features commonly identified in relation to the target variable (Chl-a) were selected, excluding nitrogen due to the sparse nature of the dataset and the absence of coexistence with other features. BOD5 was chosen over BOD as a predictor due to its relevance to the target variable in the specific water system of New York Harbor as mentioned in Chapter 4.4. The resulting dataset, after removing missing values based on row indexes of five selected features (DO, phosphorus, BOD5, temperature, and pH), consisted of 1,572 rows and 9 columns, excluding the target variable. Four additional features were included based on their relevance to Chl-a prediction in previous studies (TSS and conductivity) (Cho & Park, 2019; Z. Li et al., 2023), and the uniqueness (ammonium, and nitrate/nitrite) of the New York Harbor dataset. Subsequently, the dataset for imputations was created by using the same row indexes as the dataset of factual data, containing all 32 features, as illustrated in Table 3 (Appendix).

Each dataset was split into a 80% training and validation set and a 20% out-of-sample test set. The entire workflow was conducted within a pipeline structure. Initially, RandomForestRegressor, SVR, and MLPRegressor models from the scikit-learn library were used to build the basic models.

Hyperparameter tuning was performed using 5-fold cross-validation randomized grid search with at least 10 iterations, optimizing the selected models with the training and validation data. Preprocessing was automatically carried out within the pipeline. The search range is illustrated in Table 4, 5 and 6 (Appendix). We chose to use the randomized grid search CV over the grid search CV because the former offers a more efficient, unbiased, and flexible approach for exploring the hyperparameter space.

The tuned models were trained and evaluated using 5-fold cross-validation, with performance assessed based on average scores of evaluation metrics such as R^2, RMSE, and MAE, based on which the best performing model was chosen.

The 20% test set was used to report the final results and assess the model's generalization ability to unseen data with tuned models.

Feature permutation analysis was conducted with the best model using the permutation feature importance function from the scikit-learn library, ranking the importance of features. Hereby the whole dataset was used (training validation and test).

Based on the feature importance ranking, manual feature selection was performed to train the model with different combinations of features, and 5-fold cross-validation was used to evaluate the performance in terms of R^2, RMSE, and MAE using the training and validation set. The feature combination with the best scores will be the final choice for predictors.

6.6 *Evaluation Criteria*

The model with the highest R^2, lowest RMSE, and lowest MAE was selected from all the imputation methods, including the 'No Imputation' method. To evaluate the error pattern, scatter plots and kernel density estimation (KDE) plots were used to reveal patterns that may not be evident from the numerical results.

6.7 *Robustness*

Our models were strengthened by incorporating several techniques: 1) Consistent rows within each group ensured the validity of model comparisons among the subgroups. 2) Preprocessing steps after data splitting ensured that the statistics and transformations applied were based only on the training data and did not leak information from the test set, which helped to provide a more realistic evaluation of the model's performance on unseen data. 2) We employed 5-fold cross-validation at different stages due to its data efficiency, particularly considering our small dataset size. This approach yielded robust and reliable estimates, ensuring the generalizability of the outcomes. This technique was applied during hyperparameter tuning, training and validation, as well as feature permutation and feature selection. 3) A random state of 12 was implemented across data splitting, regressors, cross-validation, and feature permutation and feature selection to ensure reproducible results. 4) The out-of-sample approach ensured that the holdout set remained unseen until the models were tuned. 5) All operations were conducted within a pipeline. The pipeline framework prevented data leakage. In contrast, manually performing each step of the workflow may result in less organized code and require significant additional work when modifying a single preprocessing step. Given the numerous preprocessing steps involved in this project, a traditional approach without a pipeline could be time-consuming and error-prone. 6) We have chosen a randomized grid CV search over a grid search CV because the former allowed us to search a wider range of values in a much shorter time, and also reduced the risk of overfitting.

6.8 *Actual Implementation*

The actual implementation of the project utilized the following programming languages, versions, packages, and proprietary applications: Python (3.10.11), Scikit-learn (1.2.2), NumPy (1.22.4), Pandas (1.5.3), Matplotlib (3.7.1), Missingno (0.5.2), Seaborn version(0.12.2), fancyimpute (0.7.0) and Google Colaboratory (Colab).

7 RESULTS

This section presents the outcomes of optimal hyperparameters for each model, as indicated in Table 7 (Appendix). Subsequently, the model's performance will be assessed using evaluation metrics, as displayed in Table 1. Furthermore, the correspondence between predicted and actual values will be depicted through scatter plots and KDE plots, as showcased in Figures 12, 13 and 14 (Appendix). Lastly, the primary predictors identified by the chosen model will be disclosed.

7.1 *Best Hyperparameters*

The examination of Table 7 (Appendix) demonstrated that the ANN model had the same hyperparameters across various imputation methods. Conversely, the SVR model displayed varying best hyperparameters solely for the 'No Imputation' method. Similarly, the RF model had the same optimal hyperparameters for most imputation techniques, with the exception of median imputation.

Table 1: Result of Model Comparison with Different Imputation Methods. The method of using factual data is denoted as 'No Imputation'. The numbers are 5-fold cross-validation and test scores with best hyperparameters, based on R^2, RMSE and MAE. The standard errors obtained from a 5-fold cross-validation approach are reported within brackets as (s.e). The best score of R^2, RMSE and MAE of all the models across each imputation method is marked in yellow.

		R^2		RMSE		MAE	
Method	Models	Validation (s.e)	Test	Validation (s.e)	Test	Validation (s.e)	Test
No Imputation	Baseline RF	0.798 (0.011)	0.764	11.214 (0.444)	10.517	7.180 (0.180)	6.687
	SVR	0.841 (0.008)	0.770	9.956 (0.402)	10.406	6.158 (0.124)	6.167
	ANN	0.848 (0.008)	0.765	9.730 (0.379)	10.517	6.258 (0.156)	6.624
Median Imputation	Baseline RF	0.817 (0.017)	0.800	10.672 (0.723)	9.691	6.850 (0.343)	6.162
	SVR	0.843 (0.018)	0.804	9.827 (0.614)	9.595	6.150 (0.267)	5.709
	ANN	0.837 (0.018)	0.791	10.032 (0.715)	9.907	6.478 (0.291)	6.477
KNN Imputation	Baseline RF	0.798 (0.016)	0.781	11.206 (0.706)	10.154	7.145 (0.324)	6.473
	SVR	0.839 (0.015)	0.806	9.973 (0.542)	9.560	6.264 (0.253)	5.728
	ANN	0.838 (0.018)	0.774	10.000 (0.698)	10.313	6.541 (0.289)	6.549
Multiple Imputation	Baseline RF	0.820 (0.014)	0.803	10.571 (0.602)	9.637	6.717 (0.197)	5.996
	SVR	0.823 (0.018)	0.795	10.448 (0.512)	9.814	6.331 (0.217)	5.543
	ANN	0.843 (0.010)	0.773	9.876 (0.407)	10.343	6.348 (0.163)	6.620

7.2 *Baseline RF*

Among the various imputation methods, the RF model with multiple imputation demonstrated superior performance in terms of achieving the highest R^2 (0.820), as well as the lowest RMSE (10.571) and MAE (6.717) during cross-validation, as presented in Table 2. Conversely, the RF model trained solely on factual data exhibited relatively poorer performance compared to the imputed data. The model's performance remained consistent on both the validation and test sets, indicating reasonable generalization ability. The standard error of the 5-fold cross-validation was highest for median imputation and lowest for no imputation.

Figure 12 (Appendix) illustrates that a majority of data points fell within the predicted value range of 0 to 30. The 'No Imputation' method displayed slightly more scattered data points than other methods. Some data points were observed to deviate further from the prediction line around predicted values of 38 and 70, suggesting the presence of outliers.

The KDE plots revealed that the central point of the kernel density curve was slightly shifted towards the left side of 0, indicating a tendency to underestimate the target variable or the potential presence of outliers. Additionally, some bars exceeded the height of the kernel density curve at the peak region, indicating a higher density in that particular area. The imputed data, especially when using median imputation and KNN imputation, displayed a small bump on the right side of the kernel density curve, suggesting a distinct distribution in a separate cluster of data points. The peak of the kernel density curve was highest with median imputation, while it was lowest with factual data. This implies that the latter exhibited a larger dispersion of residuals, indicating a less precise fit of the model.

7.3 *SVR*

For SVR models, the median imputation produced the best performance in terms of R^2 (0.843), RMSE (9.827) and MAE (6.150) on 5-fold cross-validation among all the imputation methods. The test scores varied among the methods but not far deviated from the validation scores. Further, SVR produced lowest MAE regardless of imputation methods in both validation and test phases.

Among the different imputation methods, median imputation demonstrated the best performance for SVR models in terms of R^2 (0.843), RMSE (9.827), and MAE (6.150) during 5-fold cross-validation. The test scores did not deviate far from the validation scores. Furthermore, SVR consistently yielded the lowest MAE regardless of the imputation methods used in both the validation and test phases.

Figure 13 (Appendix) displayed more evenly scattered data points around the identity line for SVR compared to the baseline RF, particularly for predicted values exceeding 80. A few potential outliers were observed in similar regions as RF.

The KDE plots indicated that the residual distribution of SVR was also negatively skewed but less so than the baseline RF, with a longer left tail. SVR exhibited a smoother kernel density curve, suggesting that its residuals were closer to a normal distribution. Moreover, the 'No Imputation', KNN and multiple imputation methods resulted in a greater number of bars exceeding the kernel density curve compared to the median method. This aligns with the evaluation metrics, where median imputation demonstrated optimal performance among all the methods for SVR.

7.4 *ANN*

The ANN model performed similarly to SVR in terms of R^2, RMSE, and MAE when using factual data and multiple imputation method. The ANN model with factual data achieved the highest R^2 score of 0.848 and the lowest RMSE score of 9.730 during validation, surpassing all other models and methods. Its performance on the test set was slightly inferior to the validation set, but the difference remained within a reasonable range (not exceeding 0.083), indicating reasonable generalization ability.

Figure 14 (Appendix) displayed data points that appeared to be scattered closely and evenly around the identity line compared to SVR with multiple imputation. The regions where potential outliers were observed appeared to be consistent with the RF and SVR models.

Similar to the baseline RF and SVR, the KDE plots of the ANN model also exhibited bars exceeding the peak of the kernel density curve. Moreover, the left tail was less smooth when using median and KNN methods compared to SVR, suggesting that the ANN model with these imputation methods may not capture the extreme values or deviations from the central tendency as effectively as the SVR model.

Overall, in terms of evaluation metrics both SVR and ANN models demonstrated superior performance compared to the baseline RF model. Based on standard errors, the 'No imputation' method yielded the most reliable estimates, while the median imputation was less effective in capturing missing data patterns compared to KNN and multiple imputations. Error analysis indicated the presence of outliers, and similar error patterns were observed across all models and imputation methods. However, considering the evaluation metrics, SVR with median imputation emerged as the best model. Consequently, SVR with median imputation was selected for feature permutation and selection analyses.

Table 2: Result of Feature Selection. Performance comparison of the chosen model SVR with median imputation before and after feature selection. The numbers are 5-fold cross-validation and test scores based on evaluation metrics (R^2, RMSE and MAE). The standard error is denoted as (s.e).

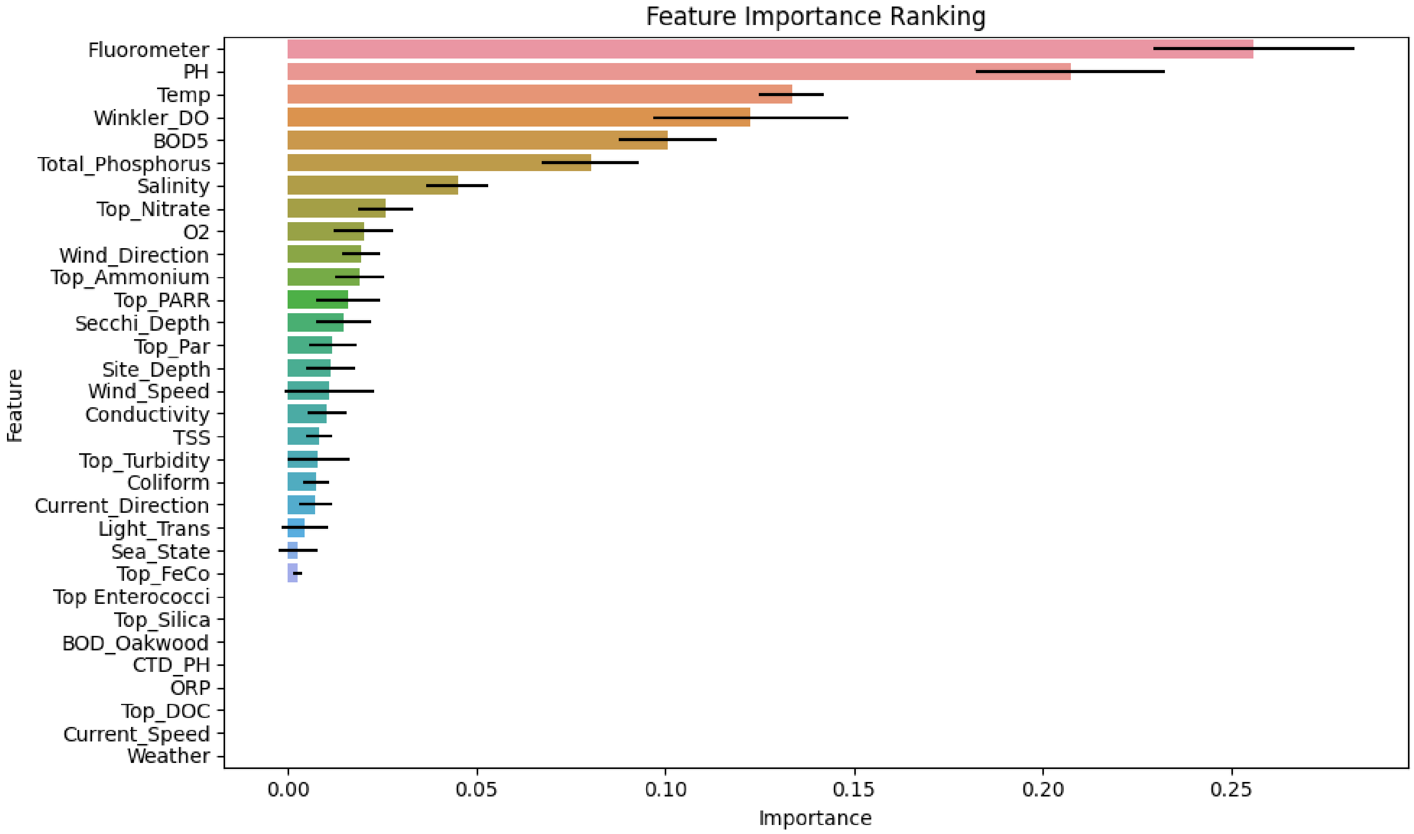
	R^2		RMSE		MAE	
	Validation (s.e)	Test	Validation (s.e)	Test	Validation (s.e)	Test
Before	0.843 (0.018)	0.804	9.827 (0.614)	9.595	6.150(0.267)	5.709
After	0.859 (0.011)	0.817	9.360 (0.512)	9.288	5.737 (0.192)	5.255

7.5 *Feature Importance Permutation and Selection*

The feature ranking depicted in Figure 4 presented all 32 features in descending order based on their importance. Upon individual feature analysis, it was observed that features with 100% missingness, such as weather, DOC, and ORP (described in Chapter 6.2.5), had no importance in predictions. This suggests that our imputation technique was effective in handling missing data up to a maximum of 30%, as evidenced by the fluorometers feature.

Out of the evaluated features, the top 17 were identified as the most important. Table 2 provided evidence of a modest enhancement in both training and testing performance resulting from this feature selection process. What's more, some of these features, namely fluorometers, nitrate/nitrite, PARR, and par, appeared to be novel in the context of HABs prediction, indicating their potential as valuable additions to the field.

Figure 4: Result of Feature Importance Ranking from Feature Permutation of SVR with Median Imputation. The length of the bar indicates the importance level of that feature over 10 iterations. The error bar is the standard deviation of the importance value. Subsequently, a feature selection analysis was performed, leading to the identification of the top 17 features that are important for HABs prediction in New York Harbor.



8 DISCUSSION

This chapter entails revisiting the problem statement and research objective, summarizing the obtained results, comparing them to relevant literature, discussing their scientific and societal impact, examining the study's limitations, and presenting suggestions for future research directions.

8.1 *Summary and Discussion of the Results*

Given the increasing prevalence of HABs and their detrimental impact on coastal regions worldwide, the timely and accurate detection of these events has become of paramount importance. The advent of machine learning techniques has offered promising avenues for addressing this imperative need. However, few studies addressed the issue of datasets with high sparsity in HABs prediction using machine learning techniques. The goal of this study is to determine whether the utilization of factual data can achieve comparable model performance in predicting HABs compared to imputed data using machine learning and deep learning algorithms, specifically, when encountering datasets containing a large amount missing values.

The dataset, obtained from New York Open Data, contained 100 columns and over 89,000 rows of water quality parameters. To answer the first sub-question SQ1, after data cleaning, two datasets containing the same rows were prepared: one with factual data and another for imputations. The performance of different models was compared, including RF as the baseline, SVR, and ANN, evaluated using R^2, RMSE, and MAE. A pipeline architecture was implemented for robust model development, involving data preprocessing, hyperparameter tuning, and validation. The best-performing model was selected based on chosen evaluation metrics. Subsequently, this model was used to perform feature importance permutation and selection, by which the second sub-question SQ2 has been answered. By combining the two sub-questions, the main research question was automatically answered in the process.

The results showed that SVR with median imputation yielded the optimal outcomes based on the evaluation metrics, while RF, the baseline model, exhibited the lowest performance. SVR demonstrated competitiveness in terms of MAE but did not outperform ANN in terms of RMSE or R^2 when factual data and multiple imputation were utilized, implying limitations in handling outliers and capturing intricate patterns under such methods. However, the performance disparities exhibited in the chosen evaluation metric and error patterns between SVR and ANN were relatively minor. Furthermore, it was noted that the identical model, regardless of the presence or absence of imputations, exhibited comparable R^2, RMSE, and MAE scores, along with uniform error patterns. Based on this observation, we can speculate that employing actual data for HABs prediction, without relying on imputation techniques, is indeed plausible.

8.2 *Comparison to the Literature*

Our experiment did not show the superiority of either machine learning algorithms or deep learning algorithms in terms of model performance, which contradicts earlier research (Azrour et al., 2022; Djerioui et al., 2019; Haghiabi et al., 2018; Koranga et al., 2022; Ooi et al., 2021; S. Zhu et al., 2019). Additionally, some literature mentioned that SVM could be computationally expensive when the data point has high dimensionality (Cruz et al., 2021; Deng et al., 2021). In our study, SVR was the fastest model, and ANN took the most time, especially during randomized grid search and cross-validations. We suspect that the reason could be that our dataset is not big enough to cause excessive computation time for SVR.

We identified the most 17 influential features, such as DO, pH, and temperature, for HABs prediction in New York Harbor. This finding reaffirmed the value of domain knowledge in feature selection. By incorporating five features extracted from previous studies (Cho & Park, 2019; Deng et al., 2021; Jeong et al., 2022; K.-M. Kim & Ahn, 2022; S. Lee & Lee, 2018; Z. Li et al., 2023; Ly et al., 2021; Ma et al., 2020; Shin et al., 2020; Wang & Xu, 2020; Xia et al., 2020; Yajima & Derot, 2017), which ranked among the top six in feature importance permutation, we were able to expedite the initial experimentation process.

Moreover, the identification of novel predictors for HABs, such as fluorometers, ammonium, and nitrate/nitrite, in our study can be attributed to the discretion of various entities responsible for water quality management. These entities have the authority to select specific parameters for measurement based on their relevance and priorities. Hence, the absence of these particular parameters in studies conducted in other water systems does not imply their nonexistence in those systems.

Notably, features such as weather and current speed, identified as influential predictors in a monsoon-like climate (Ly et al., 2021), exhibited no prediction power in our investigation of New York Harbor. We suspect that the distinctly different climate is the reason. Unfortunately, we cannot draw conclusive findings regarding their effectiveness in predicting HABs in New York Harbor due to the complete absence of these two features after the filtering process, coupled with the inadequacy of our imputation methods.

Furthermore, Yajima and Derot (2017) expressed skepticism regarding the predictive power of models with small datasets. However, our results demonstrated that the dataset size did not hinder its prediction ability, as discussed in Chapter 7. Nevertheless, we remain cautious about the generalizability of our results, considering the limited sample size of only 1572.

The strength of this research lies in the adoption of evaluation methods and models derived from previous studies in the field, which were deemed representative and practical (Azrour et al., 2022; Deng et al., 2021; Djerioui et al., 2019; Haghiabi et al., 2018; Koranga et al., 2022; S. Lee & Lee, 2018; L. Li et al., 2019; Ly et al., 2021; Ooi et al., 2021; Zamani et al., 2023; Zheng et al., 2021). Additionally, as discussed in Chapter 6.7, the utilization of a pipeline, k-fold cross-validations, and the use of a random state ensured consistency and reproducibility of our results.

8.3 *Scientific and Societal Impact*

This study contributes to the existing framework in several ways. Firstly, it demonstrated that utilizing factual data could lead to comparable model performance in predicting HABs compared to imputed data using machine learning and deep learning algorithms. This challenged the reliance on imputation techniques and suggested the feasibility of using actual data for HABs prediction.

Secondly, the study identified new water quality parameters, such as fluorometers, nitrate/nitrite, and ammonium, as crucial predictors for HABs. These findings provide novel insights and enhance our understanding of the complex dynamics involved in HABs occurrences, contributing to the scientific literature in this field.

Furthermore, the implications of this research extended beyond academia. Efficient prediction of HABs could enable stakeholders such as environmental officers and policymakers to make informed and timely decisions regarding monitoring, prevention, and mitigation strategies. These decisions are vital for maintaining water quality, safeguarding ecosystems, and protecting human health.

8.4 *Limitations and Future Directions*

This research is subject to several limitations. Firstly, potential outliers may have been present in the data based on the error analysis, but without extensive domain knowledge, it was challenging to ascertain whether extreme data points were true outliers or associated with influential events. Secondly, due to the data-driven nature of the study, certain features, such as COD, which have been recognized as crucial predictors in previous studies, were unavailable (S. Lee & Lee, 2018; Shin et al., 2020; Yajima & Derot, 2017). Furthermore, we used BOD5 in our study, the possibility of utilizing BOD to predict HABs in New York harbor was not explored. Additionally, the sample size after filtering was relatively small, comprising fewer than 2000 samples. Thirdly, Section 6.2.4 elaborated on the method employed for eliminating highly covariant features, but it was primarily effective for variables exhibiting linear correlations. Fourthly, there is scope for further improving model performance. As discussed in Section 7, the tuned ANN model had identical hyperparameters across all imputation methods, and the limited search range for hyperparameters could be one of the reasons. Moreover, we did not conduct an analysis of missing data mechanisms, such as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), so not all datasets are suitable for our method. Lastly, the employed imputation techniques may not be adequate for addressing datasets with more than 70% missingness. Notably, median imputation had higher standard error compared to other methods, indicating greater uncertainty and variability in predictions. Interestingly, the hyperparameters of ANN remained identical across all methods, suggesting its robust learning capabilities, but we could also suspect that the search range was not wide enough.

To address the limitations encountered in this project, future studies may consider the following strategies. Firstly, collaborating with domain experts would assist in effectively handling outliers and gaining deeper insights into their significance. Secondly, utilizing BOD instead of BOD5 as one of the predictors for HABs in New York harbor may unravel new insights in this field. Thirdly, adopting alternative techniques, such as partial dependence plots [1], to capture non-linear relations among the features could help identify more highly correlated features. Moreover, gaining a comprehensive understanding of the missing data mechanisms in the dataset before deciding whether to adopt complete case analysis or incorporate imputation techniques (Heymans and Twisk, 2022). What's more, there might be potential to improve model performance by expanding the range of randomized grid search. Lastly, when imputation becomes necessary, advanced methods such as deep matrix factorization (DMF) can be employed to address datasets with high levels of missing data and to mitigate the uncertainty associated with the median imputation method (Ma et al., 2020).

[1] https://christophm.github.io/interpretable-ml-book/pdp.html

9 CONCLUSION

The main research question of the research is:

RQ How to predict harmful algal bloom events in New York Harbor by using machine learning and deep learning algorithms with and without data imputations?

This question can be effectively addressed by investigating two subquestions:

- SQ1 To what extent do machine learning and deep learning models perform differently with factual data and data undergone imputation with multiple techniques?

The performance evaluation of machine learning and deep learning models on factual data demonstrated a modest disparity in terms of their predictive capabilities. However, it is worth noting that the deep learning model exhibited relatively lower computational efficiency in comparison to the machine learning models employed in this study.

- SQ2 Which are the most important water quality indicators that contribute to harmful algal bloom events in New York Harbor?

The present study has identified 17 primary water quality indicators that influence HABs within New York Harbor. These indicators, ranked in descending order of importance, include: fluorometers, pH, temperature, DO, BOD5, TP, salinity, nitrate/nitrite levels, percentage of O2 saturation, wind direction, ammonium, PARR, secchi depth, par, site depth, wind speed, and conductivity.

In conclusion, predicting HABs in New York Harbor can be achieved by utilizing 17 water quality features, including fluorometers, pH, and temperature, as inputs for machine learning models such as RF, SVR, and ANN. Constructing datasets with factual data based on domain knowledge and employing imputation techniques like median imputation, KNN imputation, and multiple imputation for up to 30% of missing values can enhance the accuracy of predictions. Our findings confirmed the feasibility of integrating factual data, machine learning algorithms, and domain knowledge to improve prediction efficiency. This approach has the potential to facilitate proactive measures by the government in mitigating the negative impacts of HABs through timely detection. Future research endeavors can capitalize on the insights and methodologies derived from our study to accelerate their investigations, particularly when faced with constraints of limited time and resources.

REFERENCES

Akhtar, N., Syakir Ishak, M. I., Bhawani, S. A., & Umar, K. (2021). Various natural and anthropogenic factors responsible for water quality degradation: A review. *Water*, *13*(19). https://doi.org/10.3390/w13192660

Anderson, D. M., Fensin, E., Gobler, C. J., Hoeglund, A. E., Hubbard, K. A., Kulis, D. M., Landsberg, J. H., Lefebvre, K. A., Provoost, P., Richlen, M. L., Smith, J. L., Solow, A. R., & Trainer, V. L. (2021). Marine harmful algal blooms (habs) in the united states: History, current status and future trends [Global Harmful Algal Bloom Status Reporting]. *Harmful Algae*, *102*, 101975. https://doi.org/10.1016/j.hal.2021.101975

Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering*, *9*(1), 104599. https://doi.org/10.1016/j.jece.2020.104599

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, *8*, 2793–2801. https://doi.org/10.1007/s40808-021-01266-6

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International journal of methods in psychiatric research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Batista, G., & Monard, M.-C. (2002). A study of k-nearest neighbour as an imputation method. *Hybrid Intelligent Systems, ser Front Artificial Intelligence Applications*, *30*, 251–260.

Blankenship, R. E. (2014). *Molecular mechanisms of photosynthesis*. John Wiley & Sons.

Burford, M. (1997). Phytoplankton dynamics in shrimp ponds. *Aquaculture Research*, *28*(5), 351–360. https://doi.org/10.1046/j.1365-2109.1997.00865.x

Burkholder, J. M. (1998). Implications of harmful microalgae and heterotrophic dinoflagellates in management of sustainable marine fisheries. *Ecological Applications*, *8*(sp1), S37–S62. https://doi.org/10.1890/1051-0761(1998)8[S37:IOHMAH]2.0.CO;2

Cariappa, M. (2004). Basic environmental technology–water supply, waste management and pollution control. *Medical Journal, Armed Forces India*, *60*(2), 206. https://doi.org/10.1016/S0377-1237(04)80128-X

Castrillo, M., & García, Á. L. (2020). Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Research*, *172*, 115490. https://doi.org/10.1016/j.watres.2020.115490

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., & Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, *171*, 115454. https://doi.org/10.1016/j.watres.2019.115454

Cho, H., & Park, H. (2019). Merged-lstm and multistep prediction of daily chlorophyll-a concentration for algal bloom forecast. *IOP Conference Series: Earth and Environmental Science*, *351*(1), 012020. https://doi.org/10.1088/1755-1315/351/1/012020

Chou, J.-S., Ho, C.-C., & Hoang, H.-S. (2018). Determining quality of water in reservoir using machine learning. *Ecological Informatics*, *44*, 57–75. https://doi.org/10.1016/j.ecoinf.2018.01.005

Cruz, R. C., Costa, P. R., Vinga, S., Krippahl, L., & Lopes, M. B. (2021). *A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination*. https://doi.org/10.3390/jmse9030283

Deng, T., Chau, K. W., & Duan, H. F. (2021). Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management*, *284*. https://doi.org/10.1016/j.jenvman.2021.112051

Derot, J., Yajima, H., & Jacquet, S. (2020). Advances in forecasting harmful algal blooms using machine learning models: A case study with planktothrix rubescens in lake geneva. *Harmful Algae*, *99*, 101906. https://doi.org/10.1016/j.hal.2020.101906

Djerioui, M., Bouamar, M., Ladjal, M., & Zerguine, A. (2019). Chlorine soft sensor based on extreme learning machine for water quality monitoring. *Arabian Journal for Science and Engineering*, *44*, 2033–2044. https://doi.org/10.1007/s13369-018-3253-8

Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Adv Neural Inform Process Syst*, *28*, 779–784. https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf

Glibert, P. M., Berdalet, E., Burford, M. A., Pitcher, G. C., Zhou, M., & Arístegui, J. (2018). Harmful algal blooms in coastal upwelling systems. *Oceanography*, *31*(2), 118–131.

Guo, J., Ma, Y., & Lee, J. H. (2021). Real-time automated identification of algal bloom species for fisheries management in subtropical coastal waters. *Journal of Hydro-environment Research*, *36*, 1–32. https://doi.org/10.1016/j.jher.2021.03.002

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

Hallegraeff, G. M. (2010). Ocean climate change, phytoplankton community responses, and harmful algal blooms: A formidable predictive challenge1. *Journal of Phycology*, *46*(2), 220–235. https://doi.org/10.1111/j.1529-8817.2010.00815.x

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Heymans, M. W., & Twisk, J. W. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology*, *151*, 185–188. https://doi.org/10.1016/j.jclinepi.2022.08.016

Hodson, T. O. (2022). Root-mean-square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in r* (2nd ed.). Springer. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Jeffrey, S., & Vesk, M. (1997). Introduction to marine phytoplankton and their pigment signatures. In *Phytoplankton pigments in oceanography* (pp. 37–84). UNESCO.

Jeong, B., Chapeta, M. R., Kim, M., Kim, J., Shin, J., & Cha, Y. (2022). Machine learning-based prediction of harmful algal blooms in water supply reservoirs. *Water Quality Research Journal*, *57*, 304–318. https://doi.org/10.2166/wqrj.2022.019

Kang, B.-K., & Park, J. (2021). Effect of input variable characteristics on the performance of an ensemble machine learning model for algal bloom prediction. *Journal of the Korean Society of Water and Wastewater*, *35*(6), 417. https://doi.org/10.11001/jksww.2021.35.6.417

Kim, H.-R., Soh, H. Y., Kwak, M.-T., & Han, S.-H. (2022). Machine learning and multiple imputation approach to predict chlorophyll-a concentration in the coastal zone of korea. *Water*, *14*(12). https://doi.org/10.3390/w14121862

Kim, K.-M., & Ahn, J.-H. (2022). Machine learning predictions of chlorophyll-a in the han river basin, korea. *Journal of Environmental Management*, *318*, 115636. https://doi.org/10.1016/j.jenvman.2022.115636

Kisi, O., Parmar, K. S., Mahdavi-Meymand, A., Adnan, R. M., Shahid, S., & Zounemat-Kermani, M. (2023). Water quality prediction of the yamuna river in india using hybrid neuro-fuzzy models. *Water*, *15*(6). https://doi.org/10.3390/w15061095

Koranga, M., Pant, P., Kumar, T., Pant, D., Bhatt, A. K., & Pant, R. (2022). Efficient water quality prediction models based on machine learning algorithms for nainital lake, uttarakhand [International Chemical Engineering Conference 2021 (100 Glorious Years of Chemical Engineering Technology)]. *Materials Today: Proceedings*, *57*, 1706–1712. https://doi.org/10.1016/j.matpr.2021.12.334

Lee, L.-H., & Lee, Y.-D. (2015). The impact of water quality on the visual and olfactory satisfaction of tourists. *Ocean Coastal Management*, *105*, 92–99. https://doi.org/10.1016/j.ocecoaman.2014.12.020

Lee, S., & Lee, D. (2018). Improved prediction of harmful algal blooms in four major south korea's rivers using deep learning models. *International Journal of Environmental Research and Public Health*, *15*(7). https://doi.org/10.3390/ijerph15071322

Li, L., Jiang, P., Xu, H., Lin, G., Guo, D., & Wu, H. (2019). Water quality prediction based on recurrent neural network and improved evidence theory: A case study of qiantang river, china. *Environmental Science and Pollution Research*, *26*, 19879–19896. https://doi.org/10.1007/s11356-019-05116-y

Li, Z., Chio, S. N., Gao, L., & Zhang, P. (2023). Assessing the algal population dynamics using multiple machine learning approaches: Application to macao reservoirs. *Journal of Environmental Management*, *334*, 117505. https://doi.org/10.1016/j.jenvman.2023.117505

Ly, Q. V., Nguyen, X. C., Lê, N. C., Truong, T. D., Hoang, T. H. T., Park, T. J., Maqbool, T., Pyo, J. C., Cho, K. H., Lee, K. S., & Hur, J. (2021). Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the han river, south korea. *Science of the Total Environment*, *797*. https://doi.org/10.1016/j.scitotenv.2021.149040

Ma, J., Ding, Y., Cheng, J. C., Jiang, F., & Xu, Z. (2020). Soft detection of 5-day bod with sparse matrix in city harbor water using deep learning techniques. *Water Research*, *170*. https://doi.org/10.1016/j.watres.2019.115350

Masó, M., & Garcés, E. (2006). Harmful microalgae blooms (hab); problematic and conditions that induce them [The Oceans and Human Health]. *Marine Pollution Bulletin*, *53*(10), 620–630. https://doi.org/10.1016/j.marpolbul.2006.08.006

Navarro, D. (2018). Learning statistics with r: A tutorial for psychology students and other beginners.

Nick, Y., Charlotte, R., Kwiatkowska, R., Beck, C., Mellon1, D., Edwards, P., Turner, J., Nicholls, P., Fearby, G., Lewis, D., Hallett, D., Bishop, T., Smith, T., Hyndford, R., Coates, L., & Turner, A. (2019). Outbreak of diarrhetic shellfish poisoning associated with consumption of mussels, united kingdom, may to june 2019. *Euro Surveill*, *24*(35), pii=1900513. https://doi.org/10.2807/1560-7917.ES.2019.24.35.1900513

Ojea, E., Ilosvay, X. E., Salgueiro-Otero, D., Rubio, I., Tidd, A. N., Caballero, S. V., Bueno-Pardo, J., Aguión, A., Barazzetta, F., & Ameneiro, J. (2023). Research priorities for seafood-dependent livelihoods under ocean climate change extreme events. *Current Opinion in Environmental Sustainability*, *61*, 101264. https://doi.org/10.1016/j.cosust.2023.101264

Ooi, K. S., Chen, Z., Poh, P. E., & Cui, J. (2021). Bod5 prediction using machine learning methods. *Water Supply*, *22*, 1168–1183. https://doi.org/10.2166/ws.2021.202

Ortenberg, E., & Telsch, B. (2003). 42 - taste and odour problems in potable water. In D. Mara & N. Horan (Eds.), *Handbook of water and wastewater microbiology* (pp. 777–793). Academic Press. https://doi.org/10.1016/B978-012470100-7/50043-1

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In P. Perner (Ed.), *Machine learning and data mining in pattern recognition* (pp. 154–168). Springer Berlin Heidelberg.

Patrick, R. (1973). Use of algae, especially diatoms, in the assessment of water quality. *ASTM Spec. Tech. Publ.*, *528*, 76–95.

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*(3), 420. https://doi.org/10.1007/s42979-021-00815-1

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Scikit-learn contributors. (2022). Sklearn.preprocessing.onehotencoder [Accessed May 2, 2023].

Shamsuddin, I. I. S., Othman, Z., & Sani, N. S. (2022). Water quality index classification based on machine learning: A case from the langat river basin model. *Water*, *14*(19). https://doi.org/10.3390/w14192939

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S. W., Lee, C. S., Kim, T. Y., Park, M. S., Park, J., & Heo, T. Y. (2020). Prediction of chlorophyll-a concentrations in the nakdong river using machine learning methods. *Water (Switzerland)*, *12*. https://doi.org/10.3390/w12061822

Summers, K. (2020). *Water quality*. IntechOpen. https://doi.org/10.5772/intechopen.77531

Wang, X., & Xu, L. (2020). Unsteady multi-element time series analysis and prediction based on spatial-temporal attention and error forecast fusion. *Future Internet*, *12*(2). https://doi.org/10.3390/fi12020034

Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., Jia, X., Yang, C., Liu, C., Ma, S., Lin, J., Wang, X., Hou, X., Zhang, K., Gao, X., Duan, P., & Qian, C. (2020). River algal blooms are well predicted by antecedent environmental conditions. *Water Research*, *185*, 116221. https://doi.org/10.1016/j.watres.2020.116221

Yajima, H., & Derot, J. (2017). Application of the random forest model for chlorophyll-a forecasts in fresh and brackish water bodies in japan, using multivariate long-term databases. *Journal of Hydroinformatics*, *20*, 206–220. https://doi.org/10.2166/hydro.2017.010

Yu, P., Gao, R., Zhang, D., & Liu, Z.-P. (2021). Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators*, *123*, 107334. https://doi.org/10.1016/j.ecolind.2020.107334

Yuan, A., Wang, B., Li, J., & Lee, J. H. (2023). A low-cost edge ai-chip-based system for real-time algae species classification and hab prediction. *Water Research*, *233*, 119727. https://doi.org/10.1016/j.watres.2023.119727

Zamani, M. G., Nikoo, M. R., Rastad, D., & Nematollahi, B. (2023). A comparative study of data-driven models for runoff, sediment, and nitrate forecasting. *Journal of Environmental Management*, *341*, 118006. https://doi.org/10.1016/j.jenvman.2023.118006

Zhai, W., Zhou, X., Man, J., Xu, Q., Jiang, Q., Yang, Z., Jiang, L., Gao, Z., Yuan, Y., & Gao, W. (2019). Prediction of water quality based on artificial neural network with grey theory. *IOP Conference Series: Earth and Environmental Science*, *295*(4), 042009. https://doi.org/10.1088/1755-1315/295/4/042009

Zheng, L., Wang, H., Liu, C., Zhang, S., Ding, A., Xie, E., Li, J., & Wang, S. (2021). Prediction of harmful algal blooms in large water bodies using the combined efdc and lstm models. *Journal of Environmental Management*, *295*, 113060. https://doi.org/10.1016/j.jenvman.2021.113060

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment Health*, *1*, 107–116. https://doi.org/10.1016/J.EEHL.2022.06.001

Zhu, S., Heddam, S., Wu, S., Dai, J., & Jia, B. (2019). Extreme learning machine-based prediction of daily water temperature for rivers. *Environmental Earth Sciences*, *78*, 202. https://doi.org/10.1007/s12665-019-8202-7

Figure 5: Boxplots. These plots show that numerous data points are positioned beyond the interquartile range (IQR) boundaries. However, these data points are not classified as outliers due to their atypical distribution patterns.

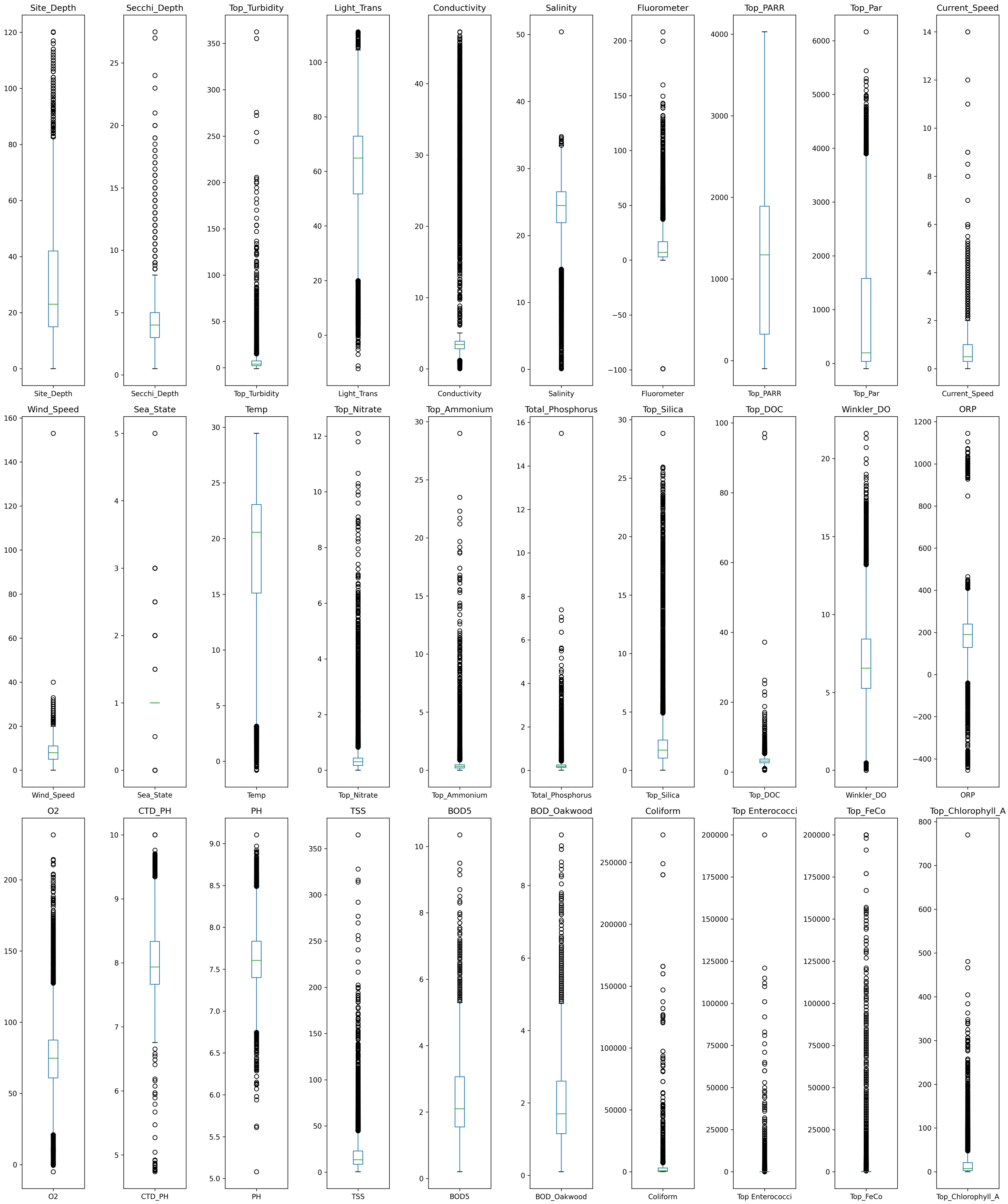


Figure 6: Data Distribution Histogram. The data distribution reveals a departure from normality for all variables, with the exception of pH and Oxygen.

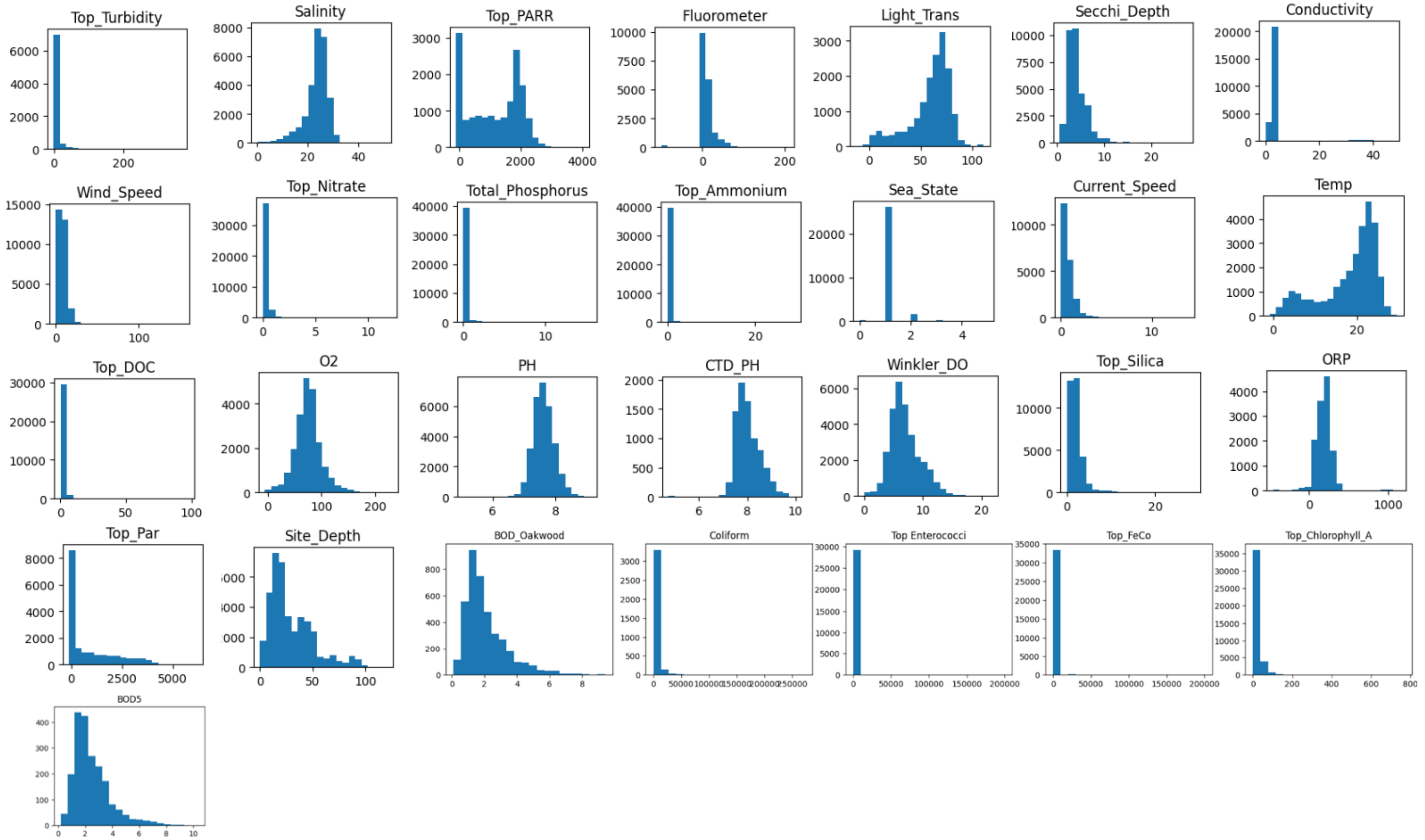


Figure 7: The Correlation Heatmap (before). It shows the linear association between variables before the removal of eight highly correlated features, with a threshold of 85% (Three categorical features are not displayed in this map).

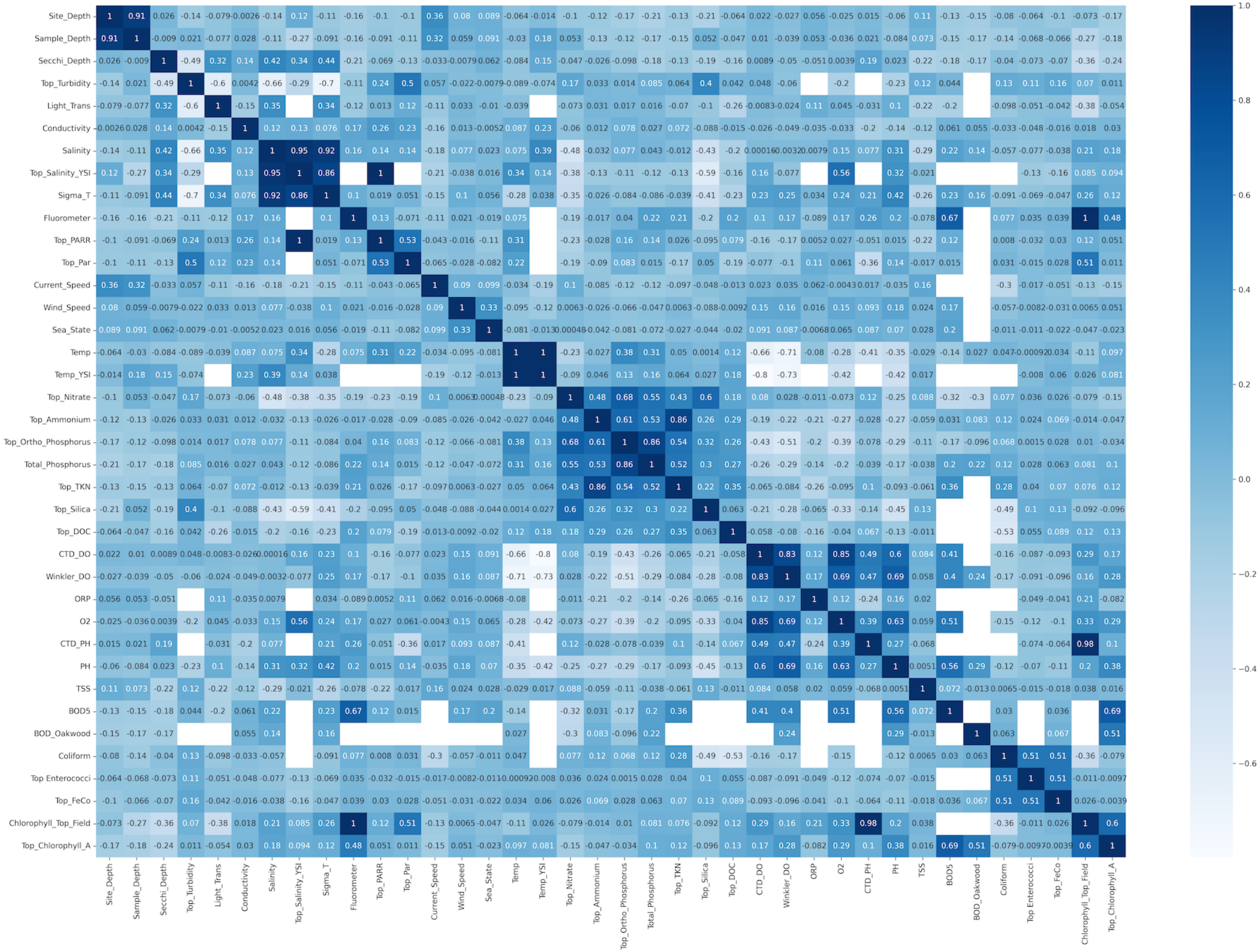


Figure 8: The Correlation Heatmap (after). It shows the linear association between variables after the removal of eight highly correlated features, with a threshold of 85% (Three categorical features are not displayed in this map).

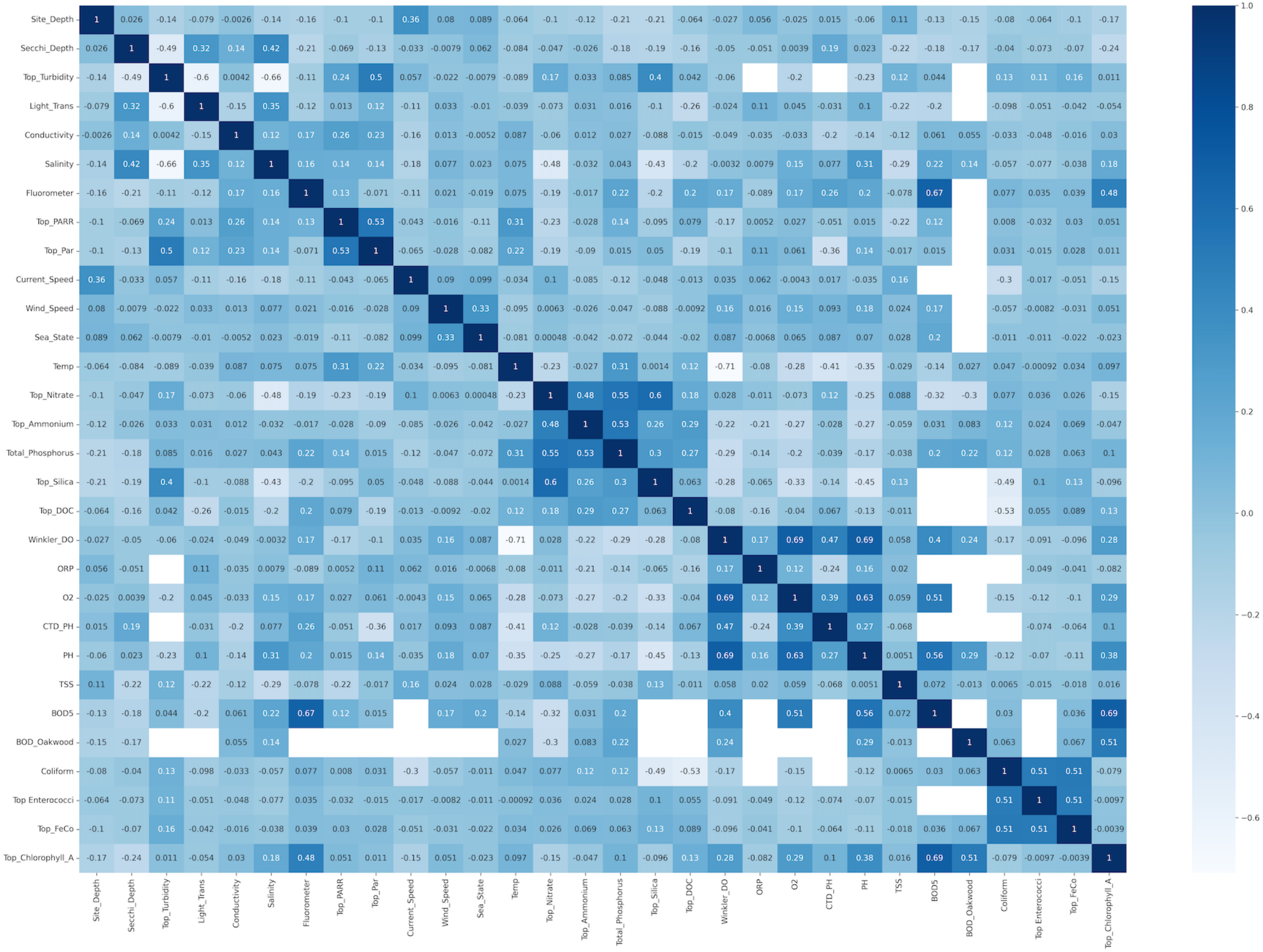


Figure 9: Raw Dataset Percentage of Missing Values. The plot depicts the percentage of missing values in the raw data, arranged in descending order. The dataset consists of 89,201 rows and 100 columns. The x-axis represents the different columns, while the y-axis represents the percentage of missing values for each feature on a scale from 0 to 1. The height of each bar corresponds to the missing percentage of the respective feature. The plot reveals that more than half of the columns contain up to 80% of missing values.

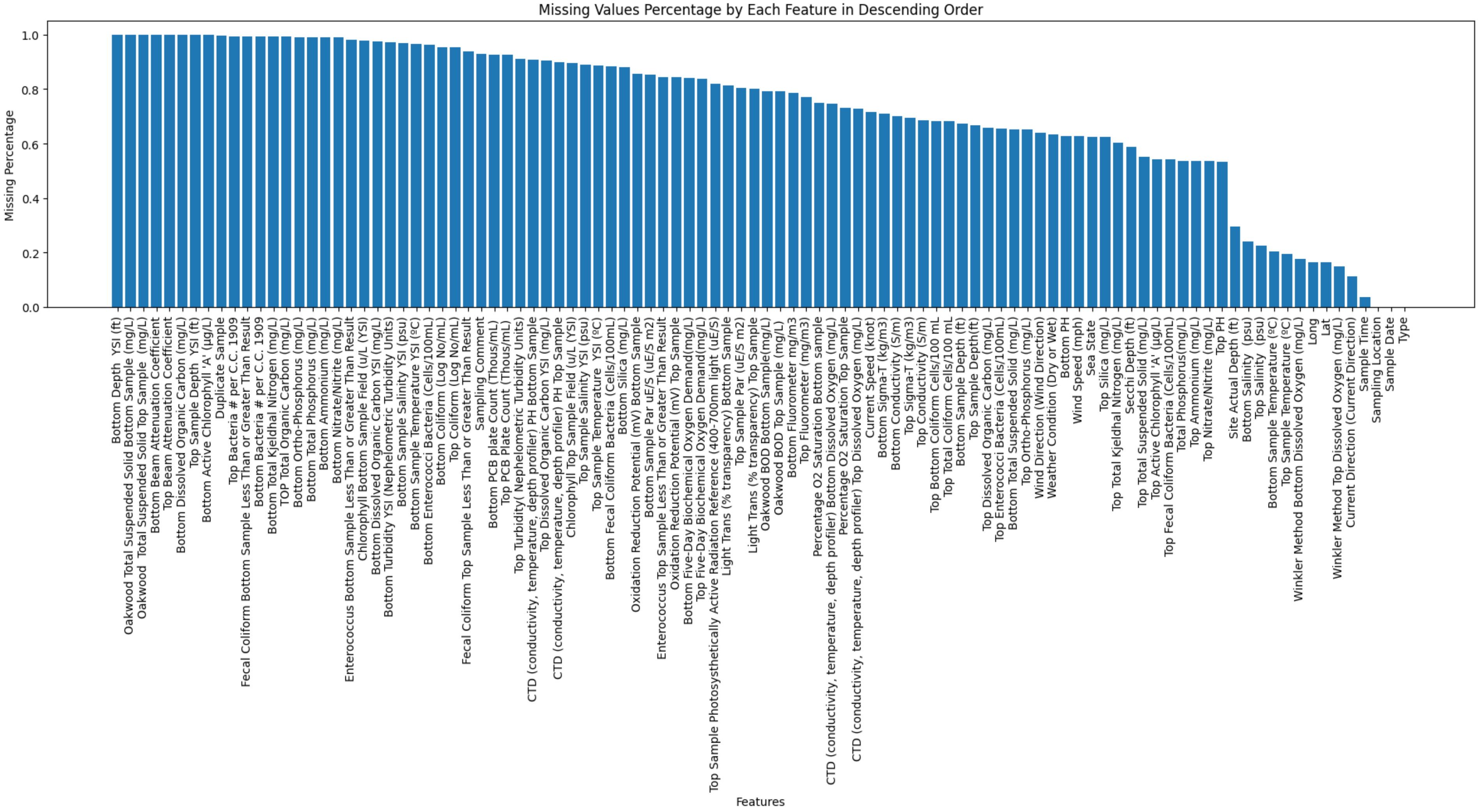


Figure 10: The correlation between categorical variables is depicted in the plots, showcasing the current direction (above) and the wind direction (below) under wet and dry weather conditions. The x-axis represents weather conditions, with 'W' denoting wet conditions and 'D' indicating dry conditions. The legend highlights each wind direction examined in Chapter 6.2. The y-axis quantifies the frequency of the analyzed categorical variables. It can be observed that the current direction consistently displays a pattern across different weather conditions, while the south, southwest, and south-southwest wind directions are more prevalent during dry weather conditions

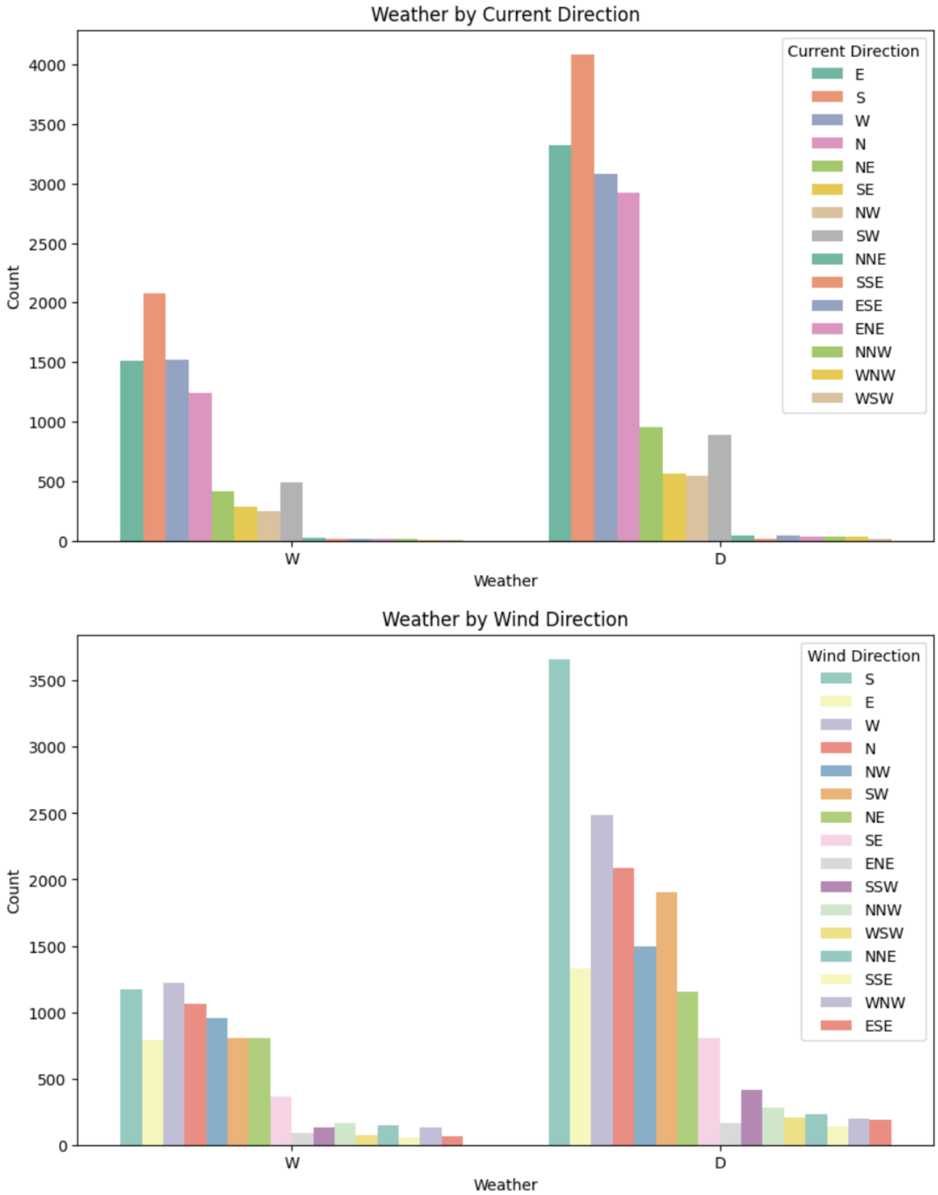


Figure 11: Entire Dataset Missing Values After Cleaning. There are 40919 rows and 33 columns after filtering out missing values of Chl-a. Many features contain a high percentage of absent data. The number on the left vertical axis indicates the percentage scale from 0 to 1. The numbers displayed on the top of the bars represent the count of non-null values for each corresponding feature.

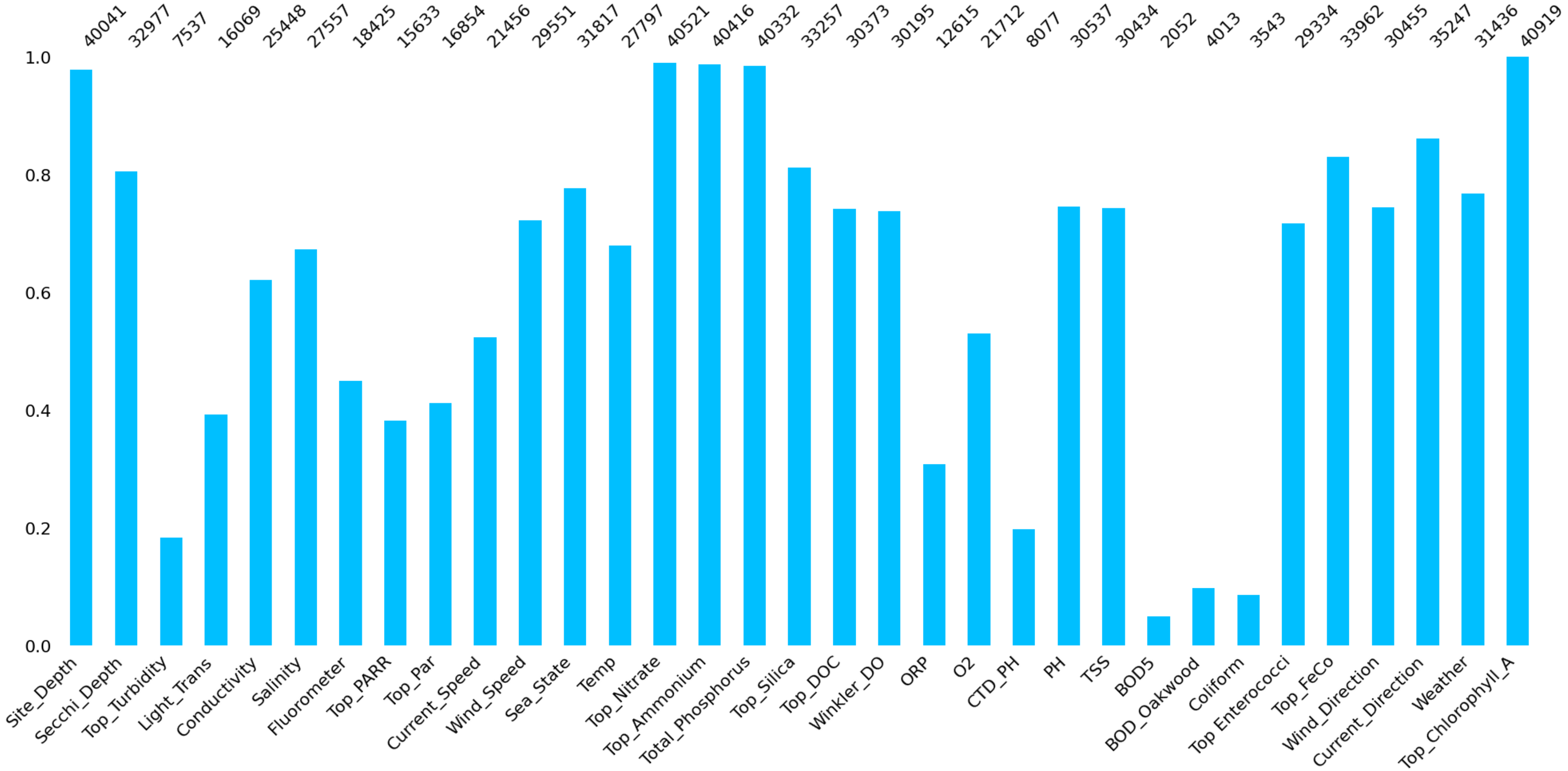


Table 3: The dataset with only factual data consists of 9 features, with the top five derived from domain knowledge, and the remaining four retained after filtering out rows with missing values. The dataset for imputations utilized the rows from the factual dataset, containing all 32 features.

Dataset with factual data	Dataset for imputations
1, Winkler DO (mg/L)	1, Winkler DO (mg/L)
2, TP (mg/L)	2, TP (mg/L)
3, BOD5 (mg/L)	3, BOD5 (mg/L)
4, Temperature (°C)	4, Temperature (°C)
5, pH	5, pH
6, Conductivity	6, Conductivity
7, Ammonium (mg/L)	7, Ammonium (mg/L)
8, TSS (mg/L)	8, TSS (mg/L)
9, Nitrate/Nitrite (mg/L)	9, Nitrate/Nitrite (mg/L)
	10, Secchi Depth (ft)
	11, Site Depth (ft)
	12, Par (uE/S m2)
	13, fluorometers (mg/m3)
	14, Salinity (psu)
	15, Silica (mg/L)
	16, DOC (mg/L)
	17, ORP (m/V)
	18, O2 Saturation
	19, Current Speed (knot)
	20, Current Direction
	21, Wind Direction
	22, Wind Speed (mph)
	23, Weather
	24, Light Trans
	25, Sea State
	26, FeCo (Cells/100 mL)
	27, Enterococci (Cells/100 mL)
	28, Turbidity
	29, CTD pH
	30, PARR (uE/S)
	31, BOD Oakwood (mg/L)
	32, Coliform (Cells/100 mL)

Nomenclature: TP, Total Phosphorus; Winkler DO, Winkler Method Top Dissolved Oxygen; TSS, Total Suspended Solid; BOD5, Five-Day Biochemical Oxygen Demand; CTD pH, CTD (conductivity, temperature, depth profiler) pH; DOC, Dissolved Organic Carbon; ORP, Oxidation Reduction Potential; O2 Saturation, Percentage O2 Saturation; Light Trans, Light Trans (% transparency); FeCo, Fecal Coliform Bacteria; PARR, Photosynthetically Active Radiation Reference.

Table 4: RF Parameter Grid for Random Search. The configuration shows the range specified for each parameter in the random search. The parameters include: 1) the number of trees, which affects the model’s complexity and the trade-off between underfitting and overfitting; 2) the maximum depth of each tree, which determines the level of interactions and complexity the model can capture; 3) the minimum number of samples required to split an internal node, which helps prevent overfitting by ensuring a minimum amount of data in each split; 4) the minimum number of samples required to be at a leaf node, which helps prevent overfitting and ensures generalization; 5) the number of features to consider when looking for the best split at each node to control the randomness and feature selection during the tree construction; and 6) the criterion that evaluates the impurity or loss during the tree splitting process.

Parameter	**Configuration**
randomforestregressor__n_estimators	randint(50, 301)
randomforestregressor__max_depth	randint(20, 51)
randomforestregressor__min_samples_split	randint(2, 21)
randomforestregressor__min_samples_leaf	randint(1, 21)
randomforestregressor__max_features	[’sqrt’, ’log2’, 1]
randomforestregressor__criterion	[’squared_error’, ’absolute_error’, ’friedman_mse’, ’poisson’]

Table 5: SVR Parameter Grid for Random Search. The configuration shows the range specified for each parameter in the random search. The parameters include: 1) C, which represents the regularization parameter that affects the complexity of the model; 2) the kernel function, which specifies the type of kernel function used in the SVR model; 3) the degree of the polynomial kernel function, which influences the complexity and capturing of non-linear interactions; 4) epsilon, which represents the width of the margin and the amount of allowed margin violations in the SVR model; 5) gamma, which specifies the kernel coefficient for kernels in the SVR model; 6) shrinking, which determines whether to use the shrinking heuristic in the SVR model, speeding up training at the cost of a potentially wider margin; and 7) tol, which sets the tolerance for stopping criteria in the SVR model's training optimization.

Parameter	**Configuration**
svr__C	uniform(0.1, 100)
svr__kernel	[′linear′, ′poly′, ′rbf′]
svr__degree	randint(0, 6)
svr__epsilon	uniform(0.001, 10)
svr__gamma	[′scale′, ′auto′]
svr__shrinking	[True, False]
svr__tol	uniform(1e-5, 1e-3)

Table 6: ANN Parameter Grid for Random Search. The configuration shows the range specified for each parameter in the random search, such as the activation function, regularization parameter (alpha), batch size, beta values, early stopping, epsilon, hidden layer sizes, learning rate, learning rate initialization, number of iterations without change before stopping, momentum, power parameter, shuffling of samples, solver algorithm, tolerance, validation fraction, and warm start option.

Parameter	**Configuration**
mlpregressor__hidden_layer_sizes	randint(50, 201)
mlpregressor__activation	[’identity’, ’logistic’, ’tanh’, ’relu’]
mlpregressor__solver	[’sgd’, ’adam’]
mlpregressor__alpha	uniform(0.0001, 1)
mlpregressor__n_iter_no_change	randint(5, 20)
mlpregressor__early_stopping	[True, False]
mlpregressor__learning_rate_init	uniform(0.0001, 0.01)
mlpregressor__shuffle	[True, False]
mlpregressor__tol	uniform(1e-5, 1e-3)
mlpregressor__warm_start	[True, False]
mlpregressor__learning_rate	[’constant’, ’invscaling’, ’adaptive’]
mlpregressor__power_t	uniform(0.1, 1.0)
mlpregressor__momentum	uniform(0.1, 1)
mlpregressor__nesterovs_momentum	[True, False]
mlpregressor__beta_1	uniform(0, 1.0)
mlpregressor__beta_2	uniform(0, 1.0)
mlpregressor__epsilon	[1e-8, 1e-7, 1e-9]
mlpregressor__validation_fraction	uniform(0.1, 1)
mlpregressor__batch_size	[’auto’, 100, 200]

Figure 12: Baseline RF Error Analysis. The first row of plots compares the predicted values to the actual values. It shows how well the model's predictions align with the true values. The red dashed line represents perfect alignment between predicted and actual values. The second row of plots displays the distribution of residuals. The x-axis represents the residuals, which are the differences between the actual target values and the corresponding predicted values. The y-axis represents the density of the residuals.

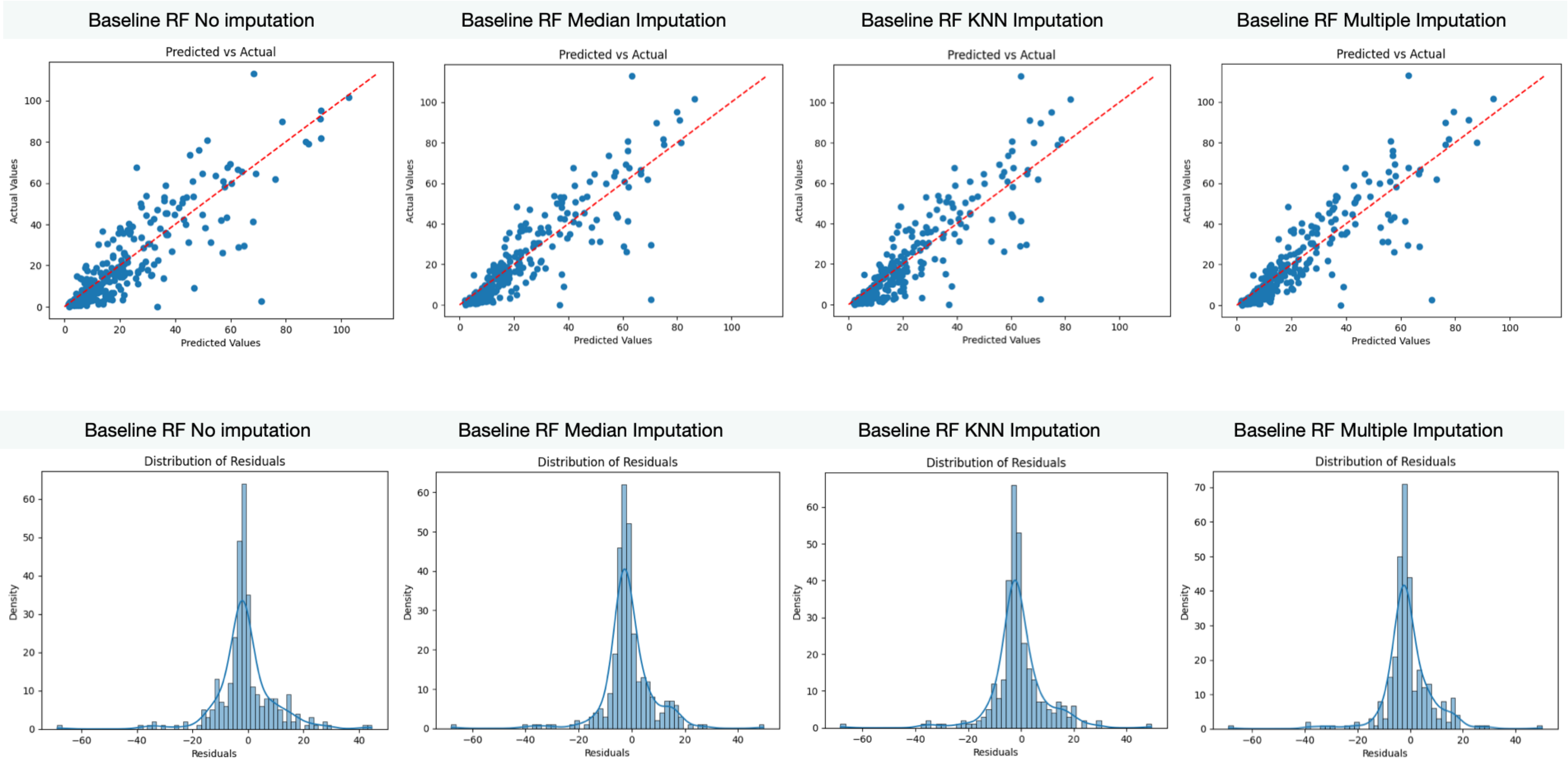
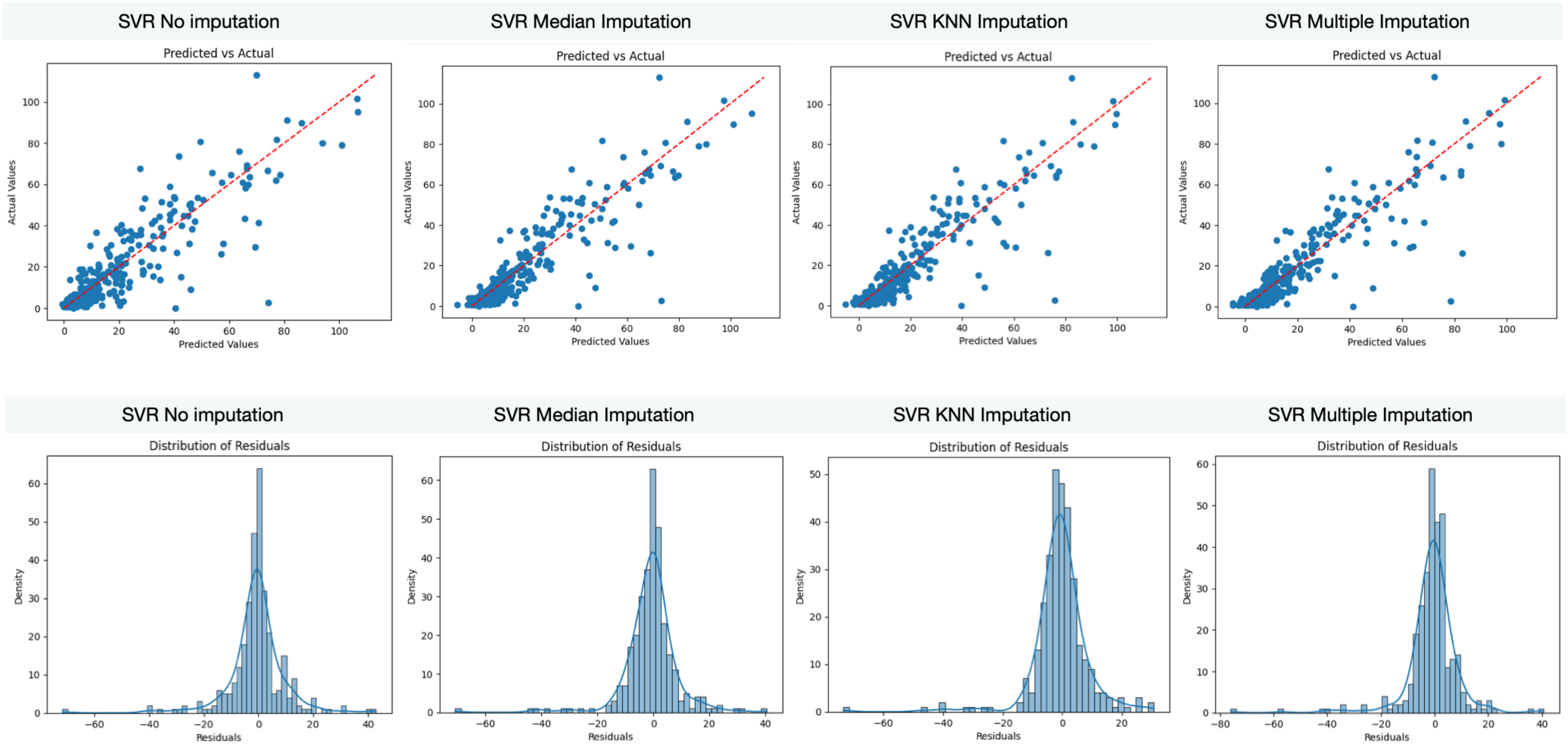


Figure 13: SVR Error Analysis. The first row of plots compares the predicted values to the actual values. It shows how well the model's predictions align with the true values. The red dashed line represents perfect alignment between predicted and actual values. The second row of plots displays the distribution of residuals. The x-axis represents the residuals, which are the differences between the actual target values and the corresponding predicted values. The y-axis represents the density of the residuals.



SVR No imputation

SVR Median Imputation

SVR KNN Imputation

SVR Multiple Imputation

Figure 14: ANN Error Analysis. The first row of plots compares the predicted values to the actual values. It shows how well the model's predictions align with the true values. The red dashed line represents perfect alignment between predicted and actual values. The second row of plots displays the distribution of residuals. The x-axis represents the residuals, which are the differences between the actual target values and the corresponding predicted values. The y-axis represents the density of the residuals.

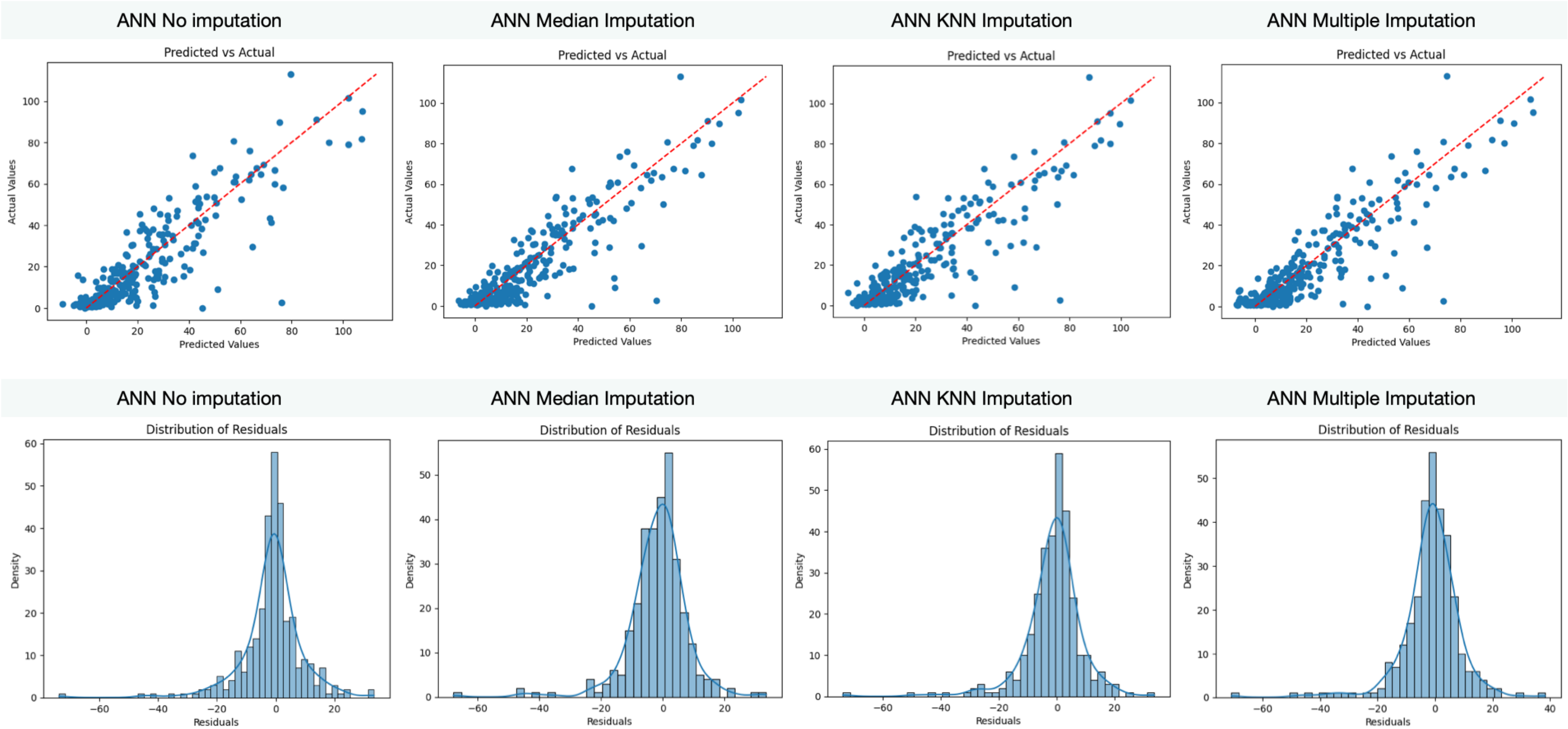


Table 7: Result of Best Hyperparameters. The values in the table correspond to the best hyperparameter settings for each model and imputation method combination. These hyperparameters were determined through a randomized grid search CV as illustrated in Chapter 6.

RF Best Parameters	**No Imputation**	**Median Imputation**	**KNN Imputation**	**Multiple Imputation**
randomforestregressor__criterion	′poisson′	′poisson′	′poisson′	′poisson′
randomforestregressor__max_depth	47	41	47	47
randomforestregressor__max_features	′sqrt′	′sqrt′	′sqrt′	′sqrt′
randomforestregressor__min_samples_leaf	3	1	3	3
randomforestregressor__min_samples_split	6	12	6	6
randomforestregressor__n_estimators	226	248	226	226
SVR Best Parameters	**No Imputation**	**Median Imputation**	**KNN Imputation**	**Multiple Imputation**
svr__C	90.171	80.227	80.227	80.227
svr__degree	5	5	5	5
svr__epsilon	0.826	1.164	1.164	1.164
svr__gamma	′auto′	′scale′	′scale′	′scale′
svr__kernel	′rbf′	′rbf′	′rbf′	′rbf′
svr__shrinking	True	False	False	False
svr__tol	0	0.001	0.001	0.001
ANN Best Parameters	**No Imputation**	**Median Imputation**	**KNN Imputation**	**Multiple Imputation**
mlpregressor__activation	′logistic′	′logistic′	′logistic′	′logistic′
mlpregressor__alpha	0.380	0.380	0.380	0.380
mlpregressor__batch_size	100	100	100	100

Continued on next page

Table 7 – continued from previous page

ANN Best Parameters	No Imputation	Median Imputation	KNN Imputation	Multiple Imputation
mlpregressor__beta_1	0.319	0.319	0.319	0.319
mlpregressor__beta_2	0.291	0.291	0.291	0.291
mlpregressor__early_stopping	False	False	False	False
mlpregressor__epsilon	1e-08	1e-08	1e-08	1e-08
mlpregressor__hidden_layer_sizes	197	197	197	197
mlpregressor__learning_rate	'adaptive'	'adaptive'	'adaptive'	'adaptive'
mlpregressor__learning_rate_init	0.009	0.009	0.009	0.009
mlpregressor__n_iter_no_change	5	5	5	5
mlpregressor__nesterovs_momentum	False	False	False	False
mlpregressor__power_t	0.796	0.796	0.796	0.796
mlpregressor__shuffle	False	False	False	False
mlpregressor__solver	'sgd'	'sgd'	'sgd'	'sgd'
mlpregressor__tol	5.758e-05	5.758e-05	5.758e-05	5.758e-05
mlpregressor__validation_fraction	0.419	0.419	0.419	0.419
mlpregressor__warm_start	True	True	True	True