Development of a Near-Infrared Spectroscopic System for Non-Invasive pH Monitoring in Water

Duy-Khanh Nguyen Viet-Hoang Ha

December 2024

Abstract

Chapter 1

Introduction

1.1 Importance of Water pH Monitoring

Water pH, a fundamental indicator of acidity or alkalinity, plays a crucial role in a multitude of natural and industrial processes involving aqueous solutions, making its accurate measurement essential across various sectors. In environmental monitoring, maintaining the appropriate pH balance in rivers, lakes, and oceans is vital for the health of aquatic ecosystems. Ocean acidification, driven by the absorption of atmospheric CO2 into seawater, continues to threaten marine organisms, particularly shellfish and coral reefs, by impairing their ability to build and maintain their shells and skeletons, with significant declines in coral reef calcification rates projected under various emission scenarios [1]. Even in freshwater systems, changes in water pH due to acid rain or industrial discharge can significantly impact the survival and reproduction of fish and other aquatic species. In industrial settings, precise control of water pH is paramount for optimizing processes and ensuring product quality. For instance, in power plants, maintaining an alkaline pH in boiler water is essential for preventing corrosion and scaling, which can reduce efficiency and lead to equipment failure [2]. In the textile industry, the pH of wastewater discharge must be carefully controlled to prevent damage to aquatic ecosystems and meet regulatory requirements [3]. In agriculture, the pH of irrigation water significantly influences nutrient availability and plant growth, with different crops requiring specific pH ranges for optimal yields. High salinity and incorrect pH of irrigation water can lead to reduced water infiltration and poor soil structure, impacting crop productivity [4]. Furthermore, in healthcare, water quality, including pH, is crucial in hemodialysis. The pH of dialysis water must be carefully controlled to prevent adverse patient reactions during treatment. Incorrect water pH can lead to acid-base imbalances in patients undergoing dialysis, highlighting the importance of stringent water quality monitoring in this critical application [5]. These examples underscore the broad and significant impact of water pH across diverse fields, highlighting the need for accurate, reliable, and preferably non-invasive pH monitoring technologies to mitigate the negative consequences of incorrect pH levels, such as environmental damage, reduced industrial efficiency, decreased agricultural yields, and compromised healthcare outcomes. The development of advanced water pH monitoring systems, such as the one proposed in this thesis, is therefore of great importance.

1.2 Limitations of Conventional pH Measurement Techniques

1.3 NIR Spectroscopy as a Solution

1.4 Review of Related Literature

1.5 Thesis Objectives

Chapter 2

Methodology

2.1 Research Design

This research presents a non-invasive method for measuring pH levels through the application of Near-Infrared (NIR) spectroscopy. This technique employs the near-infrared region of the electromagnetic spectrum to examine the molecular composition of a sample. The core principle of utilizing NIR spectroscopy for pH measurement is based on the fact that the absorption of NIR light by water molecules is affected by the pH level of the water. Variations in pH influence the hydrogen bonding network and the vibrational modes of water molecules, resulting in subtle yet quantifiable shifts in the NIR absorption spectrum. By investigating these spectral alterations, it becomes feasible to construct a predictive model that correlates the NIR spectral data with the pH value of the water sample.

Figure 2.1: Data Acquisition Process

2.2 Hardware Implementation

2.2.1 Proposed Experiment Setup

2.2.2 Hardware Components Specification

Single-board Computer

Spectroscopy Sensor

Light Source

Display Screen

Step-up Power Supply Converter Module

Step-Down Voltage Regulator

2.3 Data Acquisition Process

Fig. 2.1 specifies the sequential steps necessary for the collection and processing of spectral data from water samples, as well as the corresponding reference pH measurements.[4]This method guarantees that the data obtained is accurate,

reliable, and appropriate for further analysis and model development. The comprehensive procedure, as illustrated in the operational flowchart, is outlined below.

Each water sample was collected by adding tap water from the sink into a transparent mica box measuring 46x124x62 mm, with a thickness of 2 mm. This results in each sample containing 280 ml of water, which has a standard pH of 7.5. To change the pH of the water, 2 ml of a pH change solution, containing Bacillus licheniformis with a concentration of 1.0 x 10^6 CFU/ml, was diluted with 1.2 liters of water. This solution was then gradually added to the water samples at a rate of 1 ml at a time, allowing for a pH range adjustment from 3 to 9.

Reference pH values of the water samples were obtained using the specified pH meter that has a resolution of 0.01 and a precision of ± 0.05. Additionally, the pH meter was calibrated with a pH 4.00 buffer solution at each measurement to ensure precision.

The halogen light source was used to emit light through the water quartz to the multi-spectral sensor, which spans 18 wavelengths ranging from 410 nm to 910 nm. The setup was accurately designed and constructed using black mica, which allowed only the light from the halogen bulb to pass through. Additionally, all components were positioned to maintain a fixed and controlled arrangement, ensuring the reliability and consistency of data acquisition.

2.4 Software Implementation and Data Analysis

2.4.1 Data Preprocessing

The raw Near-Infrared (NIR) spectral data was subject to preprocessing through the application of Standard Normal Variate (SNV) transformation and Savitzky-Golay (Savgol) filtering. These methods were selected to improve the accuracy and reliability of the predictive models, taking into account the distinctive properties of NIR spectral data and the particular challenges involved in analyzing it for pH determination. NIR data presents specific difficulties compared to other spectral datasets, such as those in the visible, mid-infrared (MIR), and far-infrared (FIR) ranges, which require tailored preprocessing approaches. The NIR data is particularly sensitive to scattering effects arising from variations in particle size, sample density, and sample shape, especially in liquid samples, such as those utilized in this study. These variations can alter the optical path length and molecular absorptivity, leading to significant noise and variability within the spectra. Furthermore, due to the presence of fundamental, overtone, and combination peaks, NIR datasets frequently exhibit multicollinearity, where multiple features may correspond to the same underlying component. This characteristic can complicate the application of traditional machine learning algorithms, thereby limiting their capacity to generalize effectively.

Standard Normal Variate

The Standard Normal Variate (SNV) is a commonly utilized preprocessing technique in spectroscopy, particularly effective in reducing the impacts of light scattering and baseline variations in spectral data. Such variations typically stem from physical differences in samples, including particle size, density, and inconsistencies in the optical path length during measurement. The primary purpose of SNV is to standardize each spectrum, minimizing the effects of these physical variations and highlighting the essential chemical information vital for accurate pH assessment in this study.

Mathematically, SNV is a transformation applied to individual spectra within a dataset, consisting of two main steps: centering and scaling. Initially, the mean absorbance value of the spectrum is subtracted from each data point (centering). Subsequently, each centered data point is divided by the standard deviation of the entire spectrum (scaling). This process transforms each spectrum to have a mean of zero and a standard deviation of one.

In the context of this research, which employs NIR spectroscopy for non-invasive pH monitoring of water, SNV presents several significant advantages. The experimental setup involves directing NIR light through water samples in a cuvette, where scattering effects occur due to sample homogeneity variations and cuvette positioning. These effects can alter the light's path length, leading to inconsistent absorbance measurements. SNV mitigates the influence of these scattering variations by standardizing each spectrum.

Additionally, baseline shifts, resulting from instrumental drift, temperature changes, or other factors, can affect spectral analysis accuracy. SNV addresses these shifts by centering each spectrum around zero, ensuring that spectral differences are primarily attributed to chemical composition (pH) rather than instrumental or environmental artifacts. By minimizing the effects of physical variations and baseline shifts, SNV accentuates the subtle spectral changes associated with pH variations in water samples. This enhancement improves the capability of subsequent machine learning models (GPR and RF) to discern the relationship between NIR spectra and corresponding pH values.

Ultimately, SNV provides a cleaner, more standardized dataset, contributing to the improved performance of multivariate calibration models, resulting in more accurate and robust pH predictions, which is a key objective of this research. The SNV transformed spectrum, denoted as x_snv, is derived from a given spectrum represented as a vector x, where each element x_i corresponds to the absorbance at the wavelength i. The calculation of x_snv is conducted through the equation 2.1:

$$x_{snv,i} = \frac{x_i - \bar{x}}{s_x} \tag{2.1}$$

where:

- $x_{snv,i}$ represents the absorbance transformed through Standard Normal Variate (SNV) at wavelength i.
- x_i is the original absorbance at wavelength i.
- $\bar{x}$ is the mean absorbance of the spectrum, and s_x is the standard deviation of the spectrum.

Savitzky-Golay (Savgol) Filter

The Savitzky-Golay (Savgol) filter is a widely recognized digital signal processing method utilized for the smoothing of spectral data. Its efficacy in enhancing the signal-to-noise ratio (SNR) while minimally distorting the underlying signal is vital for precise analysis in near-infrared (NIR) spectroscopy. The filter functions by fitting a polynomial to a sliding window of data points within a spectrum, subsequently replacing the central point with the fitted polynomial value. This technique effectively averages random noise while preserving the shape and significant features of spectral peaks, which is crucial for retaining information about the chemical composition of the sample.

In the context of this research, which employs NIR spectroscopy for non-invasive pH monitoring of water, the Savgol filter presents several advantages. NIR spectra often encounter high-frequency noise from various sources, including electronic components, detector sensitivity, and fluctuations in the light source. The Savgol filter mitigates this noise, resulting in a clearer and more interpretable spectrum. Unlike basic averaging methods that may distort

spectral peaks, the Savgol filter maintains the shape, height, and width of these peaks. This preservation is critical, as subtle variations in the NIR spectral peaks of water can indicate changes in pH, making the integrity of these features essential for accurate predictions [6].

Additionally, the use of spectral derivatives (first or second) can enhance subtle features, resolve overlapping peaks, and correct baseline slopes. The Savgol filter is capable of calculating these derivatives while smoothing the data, thus enriching the information content and potentially improving the accuracy of pH prediction models. Given the inherent complexity of NIR spectra, which includes fundamental, overtone, and combination bands, as well as susceptibility to scattering effects, the Savgol filter's ability to smooth data while preserving peak characteristics renders it a highly suitable choice for this research.

The Savgol filter can be conceptualized as a weighted moving average. The smoothed spectrum, denoted as x_savgol, is obtained by applying a specific formula to each data point x_i in the original spectrum x. The output of the Savgol filter is given by the following equation:

$$x_{savgol,i} = \sum_{j=-m}^{m} C_j \cdot x_{i+j} \tag{2.2}$$

where:

- $x_{savgol,i}$ is the smoothed value at data point i.
- C_j are obtained through least-squares fitting of a polynomial of a defined order to the data points within a specified sliding window. The values of C_j are dependent upon the selected polynomial order and the size of the window.
- x_{i+j} represents the original data points within the window centered at data point i.
- m is the half-width of the window, defining the window size as $2m + 1$.

The effectiveness of the Savitzky-Golay filter relies on the careful selection of two essential parameters:

Window Size (2m+1): This parameter defines the number of data points in the sliding window. A larger window size improves smoothing but may mask sharp peaks.

Polynomial Order: This parameter denotes the degree of the polynomial used for fitting. A higher order can capture more complex curve shapes but may also be more sensitive to noise.

In this study, we will optimize the Savitzky-Golay filter parameters, specifically window size and polynomial order, to achieve an ideal balance between noise reduction and the preservation of key spectral features. The selected parameters will be recorded in the results section to ensure transparency and facilitate reproducibility of the analysis.

2.4.2 Machine Learning models

Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a non-parametric, Bayesian regression method that is particularly advantageous for modeling complex datasets, such as NIR spectral data in this study. Unlike conventional regression models that rely on predefined functional forms (e.g., linear or polynomial), GPR utilizes a prior distribution over functions, defined by a kernel function (or covariance function), which encapsulates assumptions regarding the smoothness,

periodicity, and other characteristics of the function being modeled. A Gaussian Process (GP) consists of a collection of random variables, any finite subset of which exhibits a joint Gaussian distribution. In regression, the GP establishes a distribution over functions that relate input features (NIR spectral wavelengths) to output values (pH). GPR employs this distribution to generate predictions for new, unseen input points. With a given training dataset (NIR spectra and corresponding pH values), GPR refines the prior distribution to a posterior distribution over functions, conditioned on the observed data. This posterior distribution facilitates predictions at new input wavelengths, accompanied by uncertainty estimates. GPR is particularly suited for modeling NIR spectral data and predicting pH in this project due to several key characteristics. NIR spectra frequently display complex, non-linear relationships with chemical properties like pH. The non-parametric nature of GPR enables it to capture these intricate relationships without imposing strict assumptions. The kernel function in GPR allows for the integration of prior knowledge regarding the expected smoothness and behavior of NIR spectra. Kernels such as the Rational Quadratic (RQ) or combinations of kernels (e.g., RQ + RBF) effectively model the smooth variations and multiple length scales inherent in NIR data. Additionally, GPR not only provides predictions but also quantifies the uncertainty associated with each prediction, which is essential in applications such as pH monitoring, where understanding the confidence level of predicted values is critical. The uncertainty estimates reflect the model's confidence based on the available data and the selected kernel. GPR is effective even with relatively small datasets, which is beneficial when collecting extensive NIR spectral data is challenging or costly. Furthermore, the inherently high-dimensional nature of NIR spectra, often characterized by correlated features due to overtones and combination bands, is addressed by the kernel function in GPR, making it adept at managing the multicollinearity present in NIR data. The predictive distribution for a new input point (wavelength) $x*$ in GPR is a Gaussian distribution with a mean $(x*)$ and variance $^2(x*)$. These are calculated as follows:

$$\mu(x*) = k*^T (K + \sigma_n^2 I)^{-1} y \tag{2.3}$$

where:

- $x*$ is the new input point (wavelength).
- y is the vector of observed target values (pH) in the training data.
- $k*$ is the vector of covariances between the new input point $x*$ and all training input points, computed using the kernel function: $k* = [k(x*, x), k(x*, x), ..., k(x*, x)]^T$.
- K is the kernel matrix (also called the covariance matrix), where each element $K_ij = k(x_i, x_j)$ represents the covariance between training input points x_i and x_j, calculated using the chosen kernel function.
- $k(x*, x*)$ is the covariance between the new input point and itself, computed using the kernel function.
- $_n^2$ is the noise variance, representing the assumed noise level in the observed target values.
- I is the identity matrix.

Random Forest Regressor

Random Forest (RF) is a learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is

a strong and flexible algorithm suitable for a wide range of applications, including the analysis of complex datasets like the Near-Infrared (NIR) spectral data used in this study. Unlike single decision trees, which can be prone to overfitting, Random Forests minimize this risk by assembling the predictions of multiple trees, leading to improved generalization performance and stability.

Each decision tree in a Random Forest is constructed using a bootstrap sample of the training data, meaning a random subset of the data is sampled with replacement. Additionally, at each node of the tree, only a random subset of features (wavelengths in this case) is considered for splitting. This introduces randomness into the tree-building process, separating the individual trees and reducing the variance of the overall model. During prediction, each tree in the forest independently predicts the output (pH), and these predictions are averaged to produce the final prediction.

Random Forest is particularly well-suited for modeling NIR spectral data and predicting pH in this project for several reasons. NIR spectra often exhibit complex, non-linear relationships with chemical properties such as pH. Random Forests utilize decision trees to effectively capture these interactions without requiring prior assumptions about their functional forms. Additionally, NIR spectra are high-dimensional, with each wavelength representing a unique feature. Random Forests are resilient to high dimensionality, handling feature-rich datasets while maintaining performance. Moreover, NIR spectra frequently demonstrate complex interactions among wavelengths due to overtones and combination bands. Random Forests effectively model these interactions by evaluating multiple features simultaneously during tree construction. The overall nature of Random Forests, combined with their randomization techniques, significantly reduces the risk of overfitting compared to individual decision trees. This is crucial for managing complex datasets like NIR spectra, where overfitting can be a major challenge. Lastly, by randomly selecting a subset of features at each node, Random Forests inherently perform feature selection, identifying the most relevant wavelengths for pH prediction. This is particularly beneficial in NIR spectroscopy, where many wavelengths may be excess or irrelevant.

In the context of pH prediction using Random Forest for regression, the output is determined by averaging the predictions from all individual trees within the forest. For a new input point (wavelength) x^*, the prediction can be expressed mathematically as follows:

$$\hat{y}(x*) = \frac{1}{T}\sum_{t=1}^{T} h_t(x*) \tag{2.4}$$

where:

- $\hat{y}(x*)$ is the predicted pH value for the new input point $x*$.
- T is the total number of trees in the forest
- $h_t(x*)$ is the prediction of the t^{th} decision tree for the input point $x*$.

2.4.3 Models Evaluation

R-squared (R^2)

The R-squared (R^2) score, also known as the coefficient of determination, is a statistical measure that quantifies the proportion of variance in the dependent variable — specifically pH in this context — that can be explained by

the independent variables, which consist of NIR spectral data. This score evaluates the model's goodness of fit, with values ranging from 0 to 1. A higher R^2 score indicates a better fit, with a value of 1 representing an optimal situation where the model accounts for all variability in the target variable. Conversely, an R^2 value of 0 implies that the model does not explain any variability, suggesting no improvement over a simple mean prediction.

In regression models, R^2 quantifies the alignment between the model's predictions and the actual observed values. It is crucial to recognize that while R^2 is an important metric, it does not provide insights into the accuracy of the model's assumptions or the presence of bias. A high R^2 score can sometimes be misleading, particularly in cases where the model is overfitted to the training data. Therefore, it is essential to use R^2 alongside other evaluation metrics, such as Root Mean Square Error (RMSE), and to examine residual plots to comprehensively assess the model's validity. In this research, the R^2 score will be utilized to evaluate the effectiveness of both the Gaussian Process Regression (GPR) and Random Forest (RF) models in predicting pH values based on NIR spectral data. This metric will provide a quantitative measure of how well the models capture the variance in pH and will be used in conjunction with RMSE to enable a thorough assessment of model performance, as outlined in equation 2.5:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.5}$$

where:

- SS_{res}: is the sum of squares of residuals (the difference between the observed and predicted values), calculated as $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- SS_{tot}: is the total sum of squares (the difference between the observed values and the mean of the observed values), calculated as $\sum_{i=1}^{n}(y_i - \bar{y})^2$.
- n is the number of samples
- y_i is the true value of the i^{th} sample
- $\hat{y}_i$ is the predicted value for the i^{th} sample
- $\bar{y}$ is the mean of the true values

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a widely utilized metric for assessing the discrepancies between values predicted by a model and the actual observed values. It serves to evaluate the accuracy of a model's predictions, with lower RMSE values signifying a superior alignment with the data. RMSE is particularly advantageous in scenarios where large errors are undesirable, as it assigns a relatively higher weight to significant errors due to the squaring of the differences.

The calculation of RMSE involves taking the square root of the average of the squared differences between predicted and actual values. This metric is expressed in the same units as the target variable (in this case, pH), facilitating straightforward interpretation. A lower RMSE indicates that the model's predictions are, on average, more closely aligned with the actual observed values. Within the scope of this research, a lower RMSE signifies that the predicted pH values derived from NIR spectroscopy data are in close agreement with the actual pH values measured by the reference pH meter, thereby reflecting enhanced accuracy in non-invasive pH measurement using

NIR spectroscopy. Together with the R^2 score, RMSE will be employed to conduct a thorough evaluation and comparison of the performance of the developed machine learning models, specifically Gaussian Process Regression (GPR) and Random Forest (RF), in this study. This is represented in equation 2.6:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.6}$$

where:

- n: is the number of observations.
- y_i: is the actual value for the i^{th} observation.
- $\hat{y}_i$: is the predicted value for the i^{th} observation.

2.5 System Operational flowchart

Chapter 3

Results and Discussion

3.1 Dataset Summary and Descriptive Statistics

Chapter 4

Results and Discussion

4.1 Dataset Summary and Descriptive Statistics

Given the time for the project, the research employed a dataset consisting of 500 individual water samples, each defined by 18 spectral absorbance values recorded at specific wavelengths varying from 410 nm to 940 nm, in conjunction with their respective reference pH values. The spectral data were obtained using an AS7265x spectral sensor in a controlled laboratory environment, whereas the reference pH values were determined utilizing a calibrated pH meter, which possesses a resolution of 0.01 and an accuracy of ± 0.05. A detailed description of the dataset is demonstrated in Table **??**

The dataset exhibits pH values that span from 3.08 to 9.10, thereby reflecting a comprehensive range of acidity and alkalinity levels that were assessed. The calculated mean pH is 6.14, accompanied by a standard deviation of 1.60. Furthermore, the standard deviations of the spectral absorbance values demonstrate variability across various wavelengths. Notably, the wavelengths of 435 nm, 535 nm, and 900 nm present relatively elevated standard deviations in comparison to other wavelengths, indicating a greater degree of variability in absorbance among the samples at these specific points. Conversely, the 680 nm wavelength exhibits a lower standard deviation, suggesting a reduced variability in absorbance among the samples at this wavelength.

This extensive range of pH values contributes to the development of models that are trained on a diverse dataset, thereby enhancing their potential to generalize effectively to new and previously unexamined samples. Additionally, the variability observed in the spectral data across different wavelengths underscores the critical importance of feature selection and preprocessing techniques. These methodologies are essential for identifying the most pertinent wavelengths for pH prediction and for mitigating the influence of extraneous noise. This provides a foundational understanding of the dataset's characteristics, which will be essential for the subsequent analysis, model development, and interpretation of results in the following sections.

4.2 Spectral Characteristics and Wavelength Selection Impact

4.3 Data Preprocessing Results

4.4 Machine Learning Model Performance

4.5 Performance Metrics

Chapter 5

Conclusion and Future Work

References

[1] N. R. Mollica, W. Guo, A. L. Cohen, K. F. Huang, G. L. Foster, H. K. Donald, and A. R. Solow, "Ocean acidification affects coral growth by reducing skeletal density," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1754–1759, 2018.

[2] T. A. Saleh, *Water treatment: principles and design*. John Wiley & Sons, 2021.

[3] D. A. Yaseen and M. Scholz, "Textile dye wastewater characteristics and constituents of synthetic effluents: a critical review," *International journal of environmental science and technology*, vol. 16, no. 2, pp. 1193–1226, 2019.

[4] P. S. Minhas, M. Qadir, and R. K. Yadav, "Groundwater irrigation induced soil sodification and its possible impacts on crop yields," *Agricultural water management*, vol. 228, p. 105885, 2020.

[5] R. Parvin, T. Pervin, and C. R. Ahsan, "Quality assessment of hemodialysis water in terms of physicochemical parameters and bacterial load: A study in hemodialysis units of dhaka city, bangladesh," *Journal of Environmental and Public Health*, vol. 2019, 2019.

[6] W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng, and Z. Lin, "A review of machine learning for near-infrared spectroscopy," *Sensors*, vol. 22, p. 9764, Dec. 2022.