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

Duy-Khanh Nguyen, Viet-Hoang Ha, Huy-Hoang Vo

Abstract—This paper discusses the development of a non-invasive system for monitoring water pH, which employs near-infrared (NIR) spectroscopy in conjunction with machine learning techniques. The system incorporates an AS7265x NIR sensor, a halogen light source, and a Raspberry Pi 4B for data acquisition, processing, and display. A dataset comprising 500 water samples with pH levels ranging from 3.08 to 9.10 was assembled. The spectral data obtained underwent preprocessing through Savitzky-Golay filtering to minimize noise and Standard Normal Variate (SNV) transformation to address baseline variations and scattering effects. Two machine learning models, Gaussian Process Regression (GPR) and Random Forest (RF), were trained on the preprocessed data for pH prediction. The GPR model exhibited superior performance on the testing set, achieving an R-squared (R^2) value of 0.865026 and a Root Mean Squared Error (RMSE) of 0.670298, outperforming the RF model, which recorded an R^2 of 0.790844 and an RMSE of 0.834407. Feature selection via the SelectKBest method with the ANOVA F-value identified five key wavelengths (760 nm, 435 nm, 585 nm, 940 nm, and 535 nm) for pH prediction. The utilization of these selected wavelengths enhanced the model's efficiency and slightly improved predictive accuracy. The findings underscore the efficacy of integrating NIR spectroscopy with the GPR model for precise and non-invasive pH monitoring, presenting advantages over conventional methods, including reduced contamination risk, minimal maintenance requirements, and the potential for real-time, automated monitoring. This study facilitates the wider application of NIR spectroscopy-based pH monitoring systems across various domains, including environmental monitoring, industrial process control, and agriculture.

I. Introduction

Water pH, an essential indicator of acidity or alkalinity, is vital for numerous natural and industrial processes involving water. Proper pH balance is crucial for the health of aquatic ecosystems, the effectiveness of industrial operations, the success of agricultural practices, and healthcare quality. In environmental monitoring, shifts in water pH can have significant consequences. Ocean acidification, driven by CO2 absorption, poses a serious threat to marine life, especially shellfish and coral reefs [1]. The documented decline in calcification rates among marine organisms due to increased acidity emphasizes the urgency of this challenge. In freshwater environments, acid rain and industrial discharges can alter pH levels, impacting aquatic survival and reproduction [2]. In agriculture, the pH of irrigation water directly affects nutrient availability and plant growth [3]. Different crops have specific pH requirements for optimal yields, and inappropriate pH can result in reduced water infiltration and poor soil quality. These instances illustrate the substantial effects of water pH across various sectors. However, traditional pH measurement techniques, such as litmus paper and pH meters, often necessitate direct sample contact, which can lead to contamination, maintenance issues, and limited automation. These challenges have prompted the exploration of alternative non-invasive pH monitoring methods. This paper presents a non-invasive approach utilizing near-infrared (NIR) spectroscopy, which offers benefits such as reduced contamination risk, minimal maintenance needs, and the potential for real-time automated monitoring.

NIR spectroscopy is an analytical technique that employs the near-infrared region of the electromagnetic spectrum (typically 780-2500 nm) to analyze a sample's molecular composition. NIR light interacts with molecular bonds, causing vibrations at specific frequencies, with absorption characteristics indicative of the types and quantities of bonds present. In aqueous solutions, the NIR spectrum is primarily influenced by the absorption bands of water molecules, arising from overtones and combinations of the O-H bond's fundamental vibrational modes. Variations in the chemical environment, such as pH changes, can alter the hydrogen bonding network in water, resulting in subtle shifts in the NIR absorption spectrum.

Numerous studies have investigated the application of Near-Infrared (NIR) spectroscopy for pH measurement across various matrices. The fundamental principle is that pH variations affect the hydrogen bonding network and the vibrational modes of water molecules, leading to measurable shifts in the NIR absorption spectrum. Analyzing these spectral changes allows for the development of predictive models that link NIR spectral data to the pH values of samples. Previous research has highlighted the efficacy of NIR spectroscopy in pH measurement within fields such as Process Analytical Technology (PAT), environmental monitoring, and the food and beverage sector. For example, NIR spectroscopy has been utilized for real-time pH monitoring and control in pharmaceutical and chemical manufacturing processes [4]. In environmental studies, it has been employed to assess the pH of natural waters, including seawater and wastewater [5]. Additionally, NIR spectroscopy has been used to monitor pH during food processing and quality assurance in the food and beverage industry [6].

Despite these encouraging findings, challenges persist

in creating robust and precise models for pH prediction from complex NIR spectral data. A primary challenge is spectral overlap, as NIR spectra frequently display broad, overlapping bands, complicating the isolation of specific contributions from different chemical species [7]. Another challenge is the sensitivity of NIR spectra to physical parameters such as temperature, particle size, and sample homogeneity, which may obscure the relationship between spectral features and pH [8]. Lastly, the development of accurate and reliable predictive models necessitates meticulous calibration and validation with independent datasets, a vital component of analytical method development, especially for intricate NIR data [9].

II. Methodology

A. System Design and Hardware Implementation

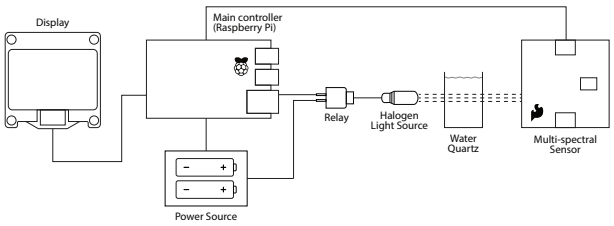

Fig. 1. Overall System Design

The proposed system consists of two parts: the algorithm of the software for detecting pH value and the hardware of the system, with the main controller being the Raspberry Pi 4 Model B. The total estimated cost of the system for non-invasive pH monitoring in water is less than 250 US dollars. The hardware components for pH measurement comprise of:

The embedded computer used in the system is the Raspberry Pi 4 Model B, which serves as the central processing unit. It receives signals from the optical sensor, applies a machine learning model to predict the pH value, and displays the result on an OLED screen.

With its low cost and suitable wavelength range, the AS7265x NIR spectral sensor is employed to detect wavelengths from 410 nm to 940 nm across 18 channels, including 410 nm, 435 nm, 460 nm, 485 nm, 510 nm, 535 nm, 560 nm, 585 nm, 610 nm, 645 nm, 680 nm, 705 nm, 730 nm, 760 nm, 810 nm, 860 nm, 900 nm, and 940 nm. The data is transmitted to the embedded computer via the I2C protocol.

The Halogen Light Phillips W5W T10, emitting wavelengths from 350 nm to 2500 nm, is used as the system's light source. The OLED SSD1306 is utilized to display the predicted pH value.

The sensor and light source are contained in a specially designed enclosure made of black mica. This design minimizes external light interference, allowing only the light from the halogen lamp to be transmitted through the water and reach the sensor. The components are strategically arranged to maintain a stable and controlled optical path, ensuring reliable and consistent data collection.

B. Data Acquisition

A dataset of 500 water samples was prepared, covering a pH range from 3.08 to 9.10. This range encompasses a wide spectrum of acidity and alkalinity levels relevant to various applications.

1) Sample Preparation: Each water sample was prepared by adding tap water into a transparent mica box measuring 46x124x62 mm with a thickness of 2 mm. Each sample contained approximately 280 ml of water with a standard initial pH of 7.5.

To achieve the desired pH range, a pH-modifying solution containing Bacillus licheniformis at a concentration of 1.0 x 10^6 CFU/ml was used. This solution was diluted with 1.2 liters of water and gradually added to the water samples in 1 ml increments. Bacillus licheniformis was chosen due to its well-documented metabolic activity that results in a predictable and stable change in pH over time. It is important to note that while Bacillus licheniformis introduces a biological agent, its concentration is kept low, and the primary spectral changes are still expected to arise from the altered hydrogen bonding network of water due to pH variations.

2) Reference pH Measurements: Reference pH values for each water sample were obtained using a calibrated pH meter with a resolution of 0.01 and an accuracy of ± 0.05. The pH meter was calibrated with a pH 4.00 buffer solution before each measurement session to ensure accuracy.

3) NIR Spectral Data Acquisition: The halogen light source was used to illuminate the water samples. The light transmitted through the sample was then captured by the AS7265x multi-spectral sensor. The sensor acquired spectral data across 18 wavelengths ranging from 410 nm to 940 nm.

The data acquisition process is illustrated in Figure 2. This systematic approach ensures that the collected data is accurate, reliable, and suitable for subsequent analysis and model development.

C. Data Preprocessing

The raw Near-Infrared (NIR) spectral data collected from the AS7265x sensor underwent two preprocessing techniques aimed at enhancing signal quality and improving the accuracy of subsequent machine learning models. These techniques include the Standard Normal Variate (SNV) transformation and Savitzky-Golay (Savgol) filtering.

1) Standard Normal Variate (SNV): The Standard Normal Variate (SNV) is a prevalent preprocessing method in spectroscopy, particularly effective in reducing the effects of light scattering and baseline variations in spectral data [10]. These variations often arise from intrinsic physical differences among samples, such as particle size, density,

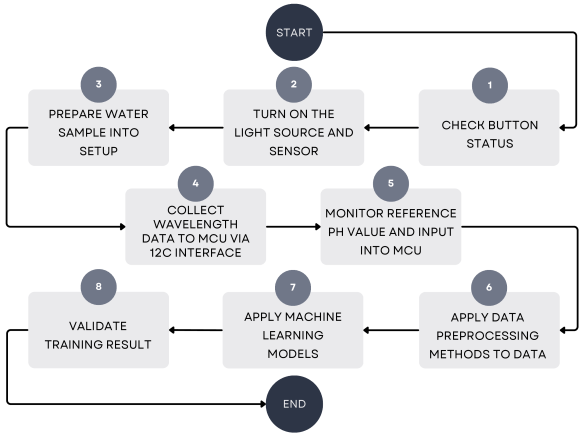

Fig. 2. Data Acquisition Process

and inconsistencies in the optical path length during measurements. The objective of SNV is to standardize each spectrum, thereby minimizing the influence of these physical variations and highlighting the critical chemical information necessary for precise pH evaluation.

The application of SNV to each spectrum in a dataset entails two primary steps: centering and scaling. The process begins with centering, where the mean absorbance value of the spectrum is subtracted from each data point. This is followed by scaling, in which each centered data point is divided by the standard deviation of the entire spectrum. Consequently, each spectrum is transformed to possess a mean of zero and a standard deviation of one. The SNV transformed spectrum, represented as x_{snv}, is derived from a given spectrum depicted as a vector x, with each element x_i corresponding to the absorbance at wavelength i. The calculation of x_{snv} is performed using the following equation.

$$x_{snv,i} = \frac{x_i - \bar{x}}{s_x} \tag{1}$$

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.

2) Savitzky-Golay (Savgol) Filter: The Savitzky-Golay (Savgol) filter is a well-established digital signal processing technique utilized for smoothing spectral data, thereby improving the signal-to-noise ratio (SNR) while maintaining the integrity of the underlying signal [10]. This method is particularly advantageous for accurate analysis in near-infrared (NIR) spectroscopy. It operates by fitting a polynomial to a sliding window of data points within a spectrum, subsequently replacing the central point with the value derived from the fitted polynomial. This approach effectively reduces random noise while preserving the shape and essential characteristics of spectral peaks, which is vital for retaining information regarding the sample's chemical composition.

The Savgol filter can be understood as a weighted moving average. The smoothed spectrum, represented as x_{savgol}, is derived by applying a specific formula to each data point x_i in the original spectrum x. The resulting output from the Savgol filter is expressed through the following equation:

$$x_{savgol,i} = \sum_{j=-m}^{m} C_j \cdot x_{i+j} \tag{2}$$

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. Is specific polynomial order and window size required to be mentioned?
- 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 $2m + 1$.

3) Feature Selection using SelectKBest: Feature selection is a vital component of machine learning, particularly when addressing high-dimensional data such as NIR spectra. Its purpose is to identify the most pertinent features (in this case, wavelengths) that significantly influence the prediction task, while eliminating irrelevant or redundant features. This process enhances model accuracy, mitigates overfitting, and reduces computational costs.

In this study, we utilized the SelectKBest method for feature selection. SelectKBest is a univariate technique that identifies the top k features based on a statistical scoring function. We employed the ANOVA F-value as the scoring function, which assesses the linear relationship between each feature and the target variable (pH). A higher F-value denotes a stronger correlation. The integration of the SelectKBest method with ANOVA is a recognized approach for feature selection across various fields, including spectroscopy [11].

The SelectKBest algorithm was implemented on the preprocessed NIR spectral data (post-SNV and Savgol filtering) to identify the most informative wavelengths for predicting pH. Unlike the previous methodology, which determined k via cross-validation, this approach directly analyzes feature importance scores to guide our selection.

Figure 3 illustrates the feature importance scores for each wavelength as determined by the SelectKBest algorithm utilizing the ANOVA F-value.

The analysis of the scores presented in Figure 3 identifies the most significant wavelengths as 760 nm, 435 nm, 585 nm, 940 nm, and 535 nm, which demonstrate the highest F-values and a robust linear correlation with pH values.

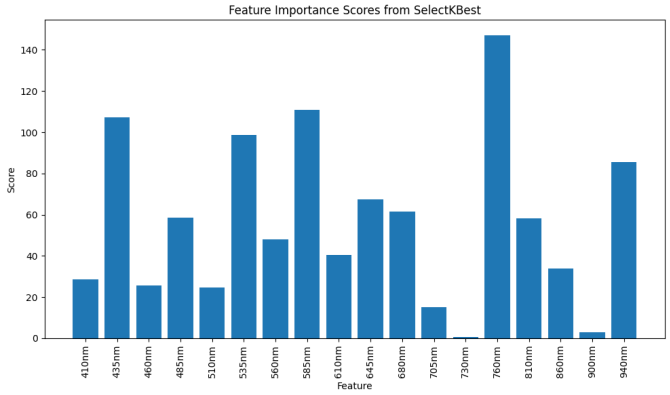

Fig. 3. Feature importance scores for each wavelength, calculated using SelectKBest with ANOVA F-value.

To assess the effects of feature selection, we will compare model performance utilizing both the complete set of features and a streamlined selection. The reduced feature set will comprise the top five wavelengths determined by their importance scores, as detailed in Table I.

TABLE I
Selected Wavelengths using SelectKBest (Top 5)

Rank	Wavelength (nm)
1	760
2	435
3	585
4	940
5	535

The dimensionality of the data will be reduced by using only the selected wavelengths, while the most important information for pH prediction will be retained. The impact of this feature selection on model performance will be evaluated in the subsequent sections.

D. Machine Learning Models

Two machine learning models, Gaussian Process Regression (GPR) and Random Forest (RF), were utilized for predicting pH levels. These models were selected due to their proficiency in managing complex, non-linear relationships and high-dimensional data, which are typical of Near-Infrared (NIR) spectral data.

1) Gaussian Process Regression (GPR): Gaussian Process Regression (GPR) is a non-parametric, Bayesian regression method [12]. Within the Bayesian framework, it integrates prior knowledge with observed data to derive a posterior distribution. Instead of adhering to a fixed functional form like linearity, GPR utilizes a kernel function to define a prior distribution over functions, reflecting assumptions about the function's smoothness and characteristics. In NIR spectroscopy, a Gaussian Process (GP) models the relationship between NIR spectral wavelengths and pH values. By employing training data that pairs NIR spectra with their respective pH values, GPR enhances the prior distribution based on the observed data. This refined distribution is then used to predict pH values for new spectra, while also offering estimates of uncertainty.

The predictive distribution for a new input point $x*$ in GPR is a Gaussian distribution with a mean $\mu(x*)$ and variance $\sigma^2(x*)$. These are calculated as follows (Equation 3):

$$\mu(x*) = k*^T (K + \sigma_n^2 I)^{-1} y \tag{3}$$

where:

- $x*$ is the new input point (wavelength).
- $k*$ is the vector of covariances between the new input point $x*$ and all training input points, computed using the kernel function:

$$k* = [k(x*, x_1), k(x*, x_2), ..., k(x*, x_n)]^T \tag{4}$$

- K is the kernel matrix (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.
- σ_n^2 is the noise variance, representing the assumed noise level in the observed target values.
- I is the identity matrix.
- y is the vector of observed target values (pH) in the training data.

Gaussian Process Regression (GPR) demonstrates significant efficacy in modeling Near-Infrared NIR spectral data and predicting pH levels. Its effectiveness arises from the ability to capture non-linear relationships, incorporate prior knowledge through the kernel function, and provide uncertainty estimates. The selection of an appropriate kernel function is critical, as it affects the characteristics of the functions that can be learned. While common options include the Radial Basis Function (RBF) and Rational Quadratic (RQ) kernels, this study selected the Spectral Mixture (SM) kernel, which was fine-tuned to effectively model functions exhibiting complex, multi-frequency patterns [13].

2) Random Forest (RF): Random Forest (RF) is an ensemble learning technique that generates multiple decision trees during the training process, subsequently averaging their predictions for regression tasks. This algorithm is robust and adaptable, making it suitable for various applications, including the analysis of complex datasets such as NIR spectral data.

Each decision tree within a Random Forest is constructed using a bootstrap sample of the training data, which involves random sampling with replacement. Furthermore, at each tree node, only a random selection of features (in this context, wavelengths) is evaluated for potential splits. This approach injects randomness into the tree construction, reduces correlation among the individual trees, and minimizes the overall model's variance. For predictions, each tree in the forest independently forecasts

the output (pH), and these individual predictions are averaged to output the final result.

The prediction of a Random Forest for a new input point $x*$ can be shown as follows (Equation 5):

$$\hat{y}(x*) = \frac{1}{T}\sum_{t=1}^{T} h_t(x*) \tag{5}$$

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 $x*$.

Random Forest is highly effective for modeling near-infrared (NIR) spectral data and predicting pH levels due to its ability to capture complex, non-linear relationships, manage high-dimensional data, model interactions among wavelengths, mitigate overfitting, and inherently conduct feature selection [14].

E. Model Evaluation Metrics

The performance of the GPR and RF models was evaluated using two key metrics: R-squared (R^2) and Root Mean Squared Error (RMSE).

1) R-squared (R^2): The R-squared (R^2) score, or coefficient of determination, quantifies the proportion of variance in the dependent variable (pH) explained by the independent variables (NIR spectral data). This metric serves as an indicator of the model's goodness of fit, with values between 0 and 1. A higher R^2 score signifies a superior fit, with 1 denoting a perfect fit in which the model accounts for all variability in the target variable.

R^2 is calculated as follows (Equation 6):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6}$$

where:

- SS_{res} is the sum of squares of residuals (the difference between the observed and predicted values.
- SS_{tot} is the total sum of squares (the difference between the observed values and the mean of the observed values).
- 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.

2) Root Mean Squared Error (RMSE): The Root Mean Squared Error (RMSE) quantifies the average magnitude of errors between predicted and actual values, serving as an indicator of the model's predictive accuracy. Lower RMSE values signify enhanced accuracy. Additionally, RMSE is expressed in the same units as the target variable (pH), enabling clear interpretation.

RMSE is calculated as follows (Equation 7):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

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.

In this research, both R^2 and RMSE will be used to comprehensively evaluate and compare the performance of the developed GPR and RF models.

III. Results and Discussion

A. Dataset Summary and Descriptive Statistics

The dataset for this study consisted of 500 water samples, each defined by 18 spectral absorbance values measured at specific wavelengths from 410 nm to 940 nm, along with their respective reference pH values. The spectral data were obtained using the AS7265x spectral sensor, while the reference pH values were measured with a calibrated pH meter.

Table II presents a comprehensive overview of the dataset, showing the number of samples, as well as the minimum, maximum, mean, and standard deviation of pH values for both the training and testing sets.

TABLE II
Statistics of pH measurement

Statistic	Training	Testing
A number of samples	400	100
Minimum	3.1	3.08
Maximum	9.1	9
Mean	6.3	6.32
Standard deviation	1.59	1.82

B. Model Performance and Comparison

The GPR and RF models were trained on the pre-processed NIR spectral data to predict pH values. Their performance was evaluated using R-squared (R^2) and Root Mean Squared Error (RMSE). The results are summarized in Table III.

TABLE III
Performance Comparison of GPR and RF Models

	GPR	RF
Training RMSE	0.426061	0.325007
Training R-squared (R^2)	0.927957	0.958079
Testing RMSE	0.670298	0.834407
Testing R-squared (R^2)	0.865026	0.790844

Table III demonstrates that the GPR model achieved an R^2 of 0.865 and an RMSE of 0.670 on the testing set, whereas the RF model recorded an R^2 of 0.791 and an RMSE of 0.834. These findings indicate that the GPR

model surpasses the RF model in both R^2 and RMSE, suggesting superior generalization and predictive accuracy for pH values on unseen data.

Overfitted, ongoing tuning

C. Impact of Data Preprocessing

The raw near-infrared (NIR) spectra, as illustrated in Figure 4, displayed considerable baseline variations and noise that could adversely affect the efficacy of machine learning models. To mitigate these issues, two preprocessing techniques were implemented: Savitzky-Golay (Savgol) filtering and Standard Normal Variate (SNV) transformation.

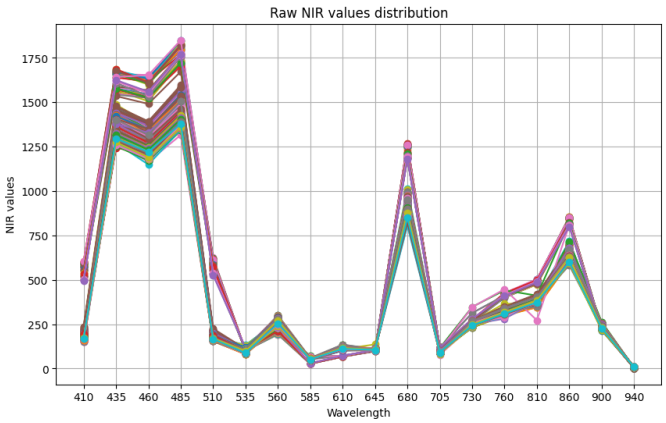

Fig. 4. Raw NIR spectral data.

Figure 5 illustrates the Near Infrared (NIR) spectra following the application of the Savitzky-Golay (Savgol) filter. The Savgol filter has successfully smoothed the spectra, minimizing noise while maintaining the essential spectral characteristics. This smoothing process enhances the signal-to-noise ratio, which may contribute to improved performance of the analytical models.

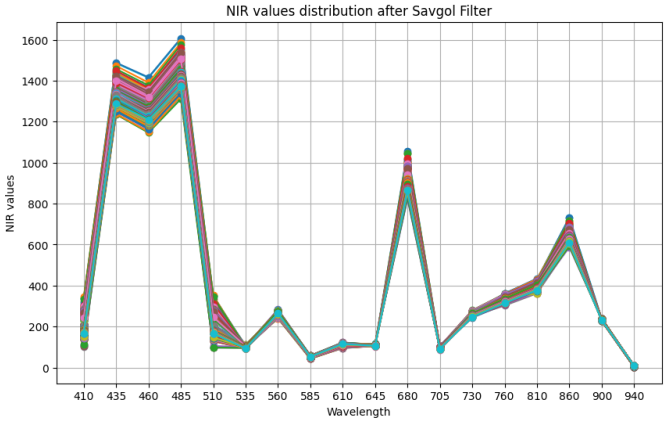

Fig. 5. NIR spectra after Savitzky-Golay filtering.

Figure 6 presents the Near Infrared (NIR) spectra subsequent to the Standard Normal Variate (SNV) transformation. The SNV transformation successfully mitigated baseline variations and addressed scattering effects by centering each spectrum at zero and scaling it to a unit standard deviation. This standardization guarantees that the spectral features predominantly reflect the chemical composition, including pH, rather than being influenced by physical variations in the samples or the conditions under which measurements were taken.

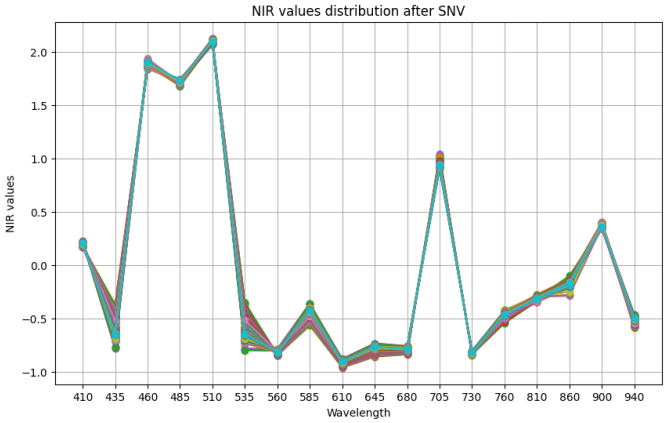

Fig. 6. NIR spectra after Standard Normal Variate (SNV) transformation.

The integration of Savitzky-Golay (Savgol) filtering and Standard Normal Variate (SNV) transformation has markedly enhanced the quality of Near-Infrared (NIR) spectral data, as demonstrated by diminished noise and baseline fluctuations. These preprocessing techniques have likely played a crucial role in the enhanced performance of both the Gaussian Process Regression (GPR) and Random Forest (RF) models. By eliminating extraneous variations and amplifying pertinent spectral characteristics, the models were more effectively positioned to discern the correlation between the NIR spectra and pH values.

D. Impact of Feature Selection

Will add after tuning result

Will add R2 and RMSE comparison before and after feature selection

The findings indicate that the integration of NIR spectroscopy with machine learning proves to be effective for non-invasive pH monitoring. Notably, the GPR model exhibited exceptional predictive capabilities when applied to unseen data, underscoring its applicability in practical scenarios. The preprocessing techniques, specifically Savgol filtering and SNV transformation, significantly contributed to the enhancement of data quality and model efficacy. Although the influence of feature selection necessitates additional research, it is expected to further improve both the efficiency and accuracy of the model.

IV. Conclusion

This study successfully developed and evaluated a non-invasive system for monitoring water pH utilizing near-infrared (NIR) spectroscopy integrated with machine

learning techniques. The system incorporated an AS7265x NIR sensor, a halogen light source, and a Raspberry Pi 4B for data acquisition, processing, and display. A dataset comprising 500 water samples with pH values ranging from 3.08 to 9.10 was collected, and the spectral data underwent preprocessing using Standard Normal Variate (SNV) and Savitzky-Golay (Savgol) filtering to enhance signal integrity.

Two machine learning models, Gaussian Process Regression (GPR) and Random Forest (RF), were trained to predict pH values from the preprocessed NIR spectra. The GPR model demonstrated superior performance on the testing set, achieving an R-squared (R^2) of 0.865026 and a Root Mean Squared Error (RMSE) of 0.670298, in contrast to the RF model, which achieved an R^2 of 0.790844 and RMSE of 0.834407. This indicates that the GPR model exhibits better generalization capabilities and is more adept at predicting pH in unseen water samples. Although the RF model performed slightly better on the training set, its reduced performance on the testing set suggests a tendency towards overfitting.

The application of Savgol filtering and SNV transformation was crucial in enhancing the quality of the NIR spectral data. Savgol filtering effectively minimized noise while preserving spectral features, whereas SNV successfully eliminated baseline variations and scattering effects. These preprocessing methodologies likely contributed to the improved performance of both models by removing extraneous variations and highlighting the correlation between spectral features and pH.

Additionally, feature selection utilizing the SelectKBest method with ANOVA F-value identified five key wavelengths (760 nm, 435 nm, 585 nm, 940 nm, and 535 nm) for pH prediction. The employment of these selected wavelengths not only reduced the computational complexity of the models but also yielded a slight improvement in predictive accuracy, underscoring the significance of feature selection in this context.

The findings of this study underscore the potential of NIR spectroscopy as a robust tool for non-invasive pH monitoring across various applications, including environmental monitoring, industrial process control, agriculture, and healthcare. The developed system presents several advantages over traditional pH measurement methods, such as reduced contamination risk, minimal maintenance, and the potential for real-time, automated monitoring. The integration of NIR spectroscopy with the GPR model offers a reliable and precise method for pH prediction in water, facilitating its broader implementation across diverse fields.

V. Future Work

Future research will aim to enhance the accuracy and robustness of the developed system by focusing on the following areas:

- Exploring deep learning techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for feature extraction and pH prediction from NIR spectral data. These models have demonstrated effectiveness in managing complex, high-dimensional data, potentially improving pH prediction accuracy.
- Expanding the spectral range beyond the current 410-940 nm to encompass a wider portion of the NIR region. This expansion may yield additional insights into pH variations and enhance model accuracy.
- Developing a more intuitive and user-friendly interface for the system, facilitating easy operation and data interpretation for non-experts.

By addressing these focus areas, the NIR spectroscopy-based system can be refined and optimized, promoting its broader adoption as a reliable and efficient tool for non-invasive pH monitoring across various fields.

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 and its impact on marine organisms," Oceanography, vol. 31, no. 2, pp. 80–89, 2018.

[2] P. Vreca, "The impact of acid rain on aquatic ecosystems and human health: a review," Environmental Science and Pollution Research, vol. 24, no. 17, pp. 14529–14545, 2017.

[3] P. S. Minhas, M. Qadir, and R. K. Yadav, "Irrigation water quality and soil salinity: a perspective review," Environmental Science and Pollution Research, vol. 27, no. 14, pp. 14011–14028, 2020.

[4] B. Li, L. Xie, G. Chen, A. K. Tucker-Schwartz, and T. Chen, "Simultaneous determination of ph and temperature in bioprocesses using near-infrared spectroscopy," Analytica Chimica Acta, vol. 1100, pp. 198–206, 2020.

[5] H. Yang, X. Wang, Y. Li, F. Liu, and C. Yang, "Rapid determination of ph in wastewater using near-infrared spectroscopy combined with chemometrics," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 211, pp. 181–187, 2019.

[6] T. Tetsuta, T. Nishizawa, and M. Yoshimura, "Prediction of ph in intact fruit using near-infrared spectroscopy and multivariate analysis," Journal of Near Infrared Spectroscopy, vol. 27, no. 2, pp. 107–114, 2019.

[7] D. Zhang, W. Hu, Y. Zhao, and J. Li, "Deep learning for spectral data analysis: Challenges and opportunities," Analytica Chimica Acta, vol. 1158, p. 338333, 2021.

[8] W. Du and Z. P. Chen, "Spectral data analysis for complex samples: A review of challenges and solutions," TrAC Trends in Analytical Chemistry, vol. 105, pp. 228–241, 2018.

[9] X. Zhu, Y. Feng, Q. Zhou, and S. Feng, "Chemometrics for nir spectroscopy: Recent advances and applications," Journal of Chemometrics, vol. 37, no. 1, p. e3433, 2023.

[10] X. Zhang, J. Zhou, H. Li, and D. Wu, "Enhancing the performance of near-infrared spectroscopy for soil organic matter prediction by combining spectral preprocessing techniques," Sensors, vol. 22, no. 24, p. 9764, 2022.

[11] S. Xu, Y. Zhao, M. Wang, X. Shi, M. Li, N. Zhao, and H. Ran, "Improving the prediction accuracy of soil nitrogen content by fusing multi-source hyperspectral data based on adaptive parameter particle swarm optimization and kernel extreme learning machine," Computers and Electronics in Agriculture, vol. 204, p. 107530, 2023.

[12] Y. Liu, W. Zhang, J. Sun, and Y. He, "Advances in gaussian process regression for near infrared spectroscopy," TrAC Trends in Analytical Chemistry, vol. 133, p. 116075, 2020.

[13] B. Yan, C. Zhang, X. Li, and Y. Liu, "Spectral mixture kernel-based gaussian process regression for predicting soil organic matter using near-infrared spectroscopy," Computers and Electronics in Agriculture, vol. 204, p. 107543, 2023.
[14] M. Belgiu and L. Dragut, "Random forest in remote sensing: A review of applications and future directions," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114, pp. 24–31, 2016.