Journal of Information Systems Engineering and Management
2025, 10(17s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Smartphone Camera-Based Prediction of Water pH Using Multispectral Image Simulation and Machine Learning Models

M Allam Daffa Alhaqi[1], Bernadetha Grace Wisdayanti[2], Chatchawan Chaichana[3*], Wahyu Nurkholis Hadi Syahputra[4], Nappasawan Womongkol[5], Suwimon Wicharuck[6]

[1,2] *Graduate Master's Degree Program in Agricultural Engineering, Department of Mechanical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand.*

[3,4,5] *Department of Mechanical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand.*

[6] *Office of Research Administration, Chiang Mai University, Chiang Mai 50200, Thailand.*

**Corresponding Author: chatchawan.c@cmu.ac.th*

ARTICLE INFO

Received: 09 Dec 2024

Revised: 29 Jan 2025

Accepted: 10 Feb 2025

ABSTRACT

This study presents an evaluation of several model for predicting water pH using smartphone cameras with different resolutions (13 MP and 48 MP). The prediction models were used from numerical data extracted from images processed to simulate multispectral properties. Water samples were prepared by adjusting tap water pH using a 5% nitric acid (HNO3) and 10% potassium hydroxide (KOH), covering a range of pH values 4 - 8. Each sample was captured using both 13 MP and 48 MP smartphone cameras, and digital filtering over the visible spectrum (380-780 nm) was applied through image processing. From these images, the value of Red, Green, Blue, and Grayscale values were extracted, along with derived metrics as well as luminance, hue, saturation, and coloration index. These features were used as inputs for machine learning models to predict pH value. Among the algorithms tested, the Decision Tree (DT) model achieved the highest prediction accuracy, outperforming Random Forest (RF) and Multi-Layer Perceptron Neural Network (MLP-NN), while 48 MP camera outperforming 13 MP camera. These findings show the feasibility of predicting water pH from smartphone-captured images processed to simulate multispectral characteristics. Further work with an expanded dataset may enhance the model's accuracy and precision.

Keywords: Digital image processing, Multispectral simulation, Machine learning, Model development, pH Prediction.

INTRODUCTION

Water is a valuable natural resource, particularly crucial in agriculture, where it supports crop irrigation systems. The quality of irrigation water directly impacts crop yields and plant health, making it essential to monitor and manage water quality [1], [2]. One important aspect of water quality is pH, which influences both crop health and soil conditions. Monitoring and adjusting water pH to meet specific plant needs has become an essential practice in agriculture, as it ensures optimal water quality for plant growth and productivity. Periodic and rapid assessments of water pH are therefore critical in agricultural irrigation, helping farmers maintain suitable conditions for diverse crop

requirements.

Recently, various methods for assessing water quality have been reported, each with unique advantages and limitations. Laboratory testing remains the standard due to its high accuracy, but it requires specialized equipment and chemical reagents, making it expensive and time-consuming [3]. Commercial water test kits, particularly paper-based instruments, are also widely used for on-site water quality assessments. These kits rely on color changes by chemical reactions, which can be interpreted visually or with digital tools. However, to reduce potential errors in visual interpretation, digital image processing and neural network approaches have been explored to improve accuracy and consistency in parameter prediction. Although these methods provide more reliable results, paper-based test kits still require single-use strips for each measurement, which can be costly for frequent testing [4].

Spectrometer devices offer another approach for water quality assessment. These optical sensors measure radiant energy across specific wavelengths, providing detailed spectral information on water samples. For instance, spectrometers often operate within a specific spectral range, and users can select certain wavelengths or intervals within that range to focus on certain characteristics of the water sample, depending on the observational requirements [5]. Recently, Li proposed a novel remote sensing method for water quality assessment using a modified micro-hyper spectrometer [6]. This system provides long-term durability and autonomous monitoring capabilities, making it ideal for continuous water quality monitoring. However, the device's large size and high cost, combined with its complex components, limit its portability and affordability. As a result, there is a growing demand for a low-cost, handheld optical sensing system that is easy to use and accessible to small-scale farmers, allowing for rapid and convenient water quality assessment in agricultural irrigation applications.

Smartphone-based analysis is appealing due to its accessibility, portability, and capability to capture high-resolution images with recent advancements in smartphone cameras. These technology holds potential for provide rapid and low-cost pH analysis for agricultural and environmental purposes. The integration of mobile technology with innovative sensing techniques has shown promising results in accurately measuring these parameters. While mobile camera-based water parameter assessment offers numerous advantages, including cost-effectiveness and accessibility, it also faces challenges such as ensuring data accuracy and dealing with environmental variability. Continued advancements in sensor technology and data processing algorithms are essential to overcome these challenges and enhance the reliability of mobile-based water quality monitoring systems.

Gozukara conducted a study on predicting soil electrical conductivity (EC) using a smartphone-based system (iPhone 11) by analyzing color coordinates such as RGB, HSV, and CIE Lab* [7]. Their findings indicated that smartphone and visible spectrum (Vis)-based color coordinates could be used to predict soil EC and categorize salinity levels. Using a Random Forest (RF) algorithm with RGB values and various indices (Brightness, Saturation, Hue, Coloration, and Redness), they achieved an R2 value of 0.51, showing moderate predictive accuracy. While this study demonstrated the potential of smartphone-based methods for soil EC prediction, the moderate R2 value suggests limitations in prediction accuracy, especially for precise applications. These results highlight the need for further improvements in image processing, feature extraction, and model refinement to enhance prediction accuracy. Additionally, further research is needed to explore how well these models generalize across different environmental conditions and devices, particularly for applications in water quality prediction where factors and requirements differ.

This study demonstrates the potential of developing and evaluating a machine learning model to predict water pH

using datasets extracted from images captured by smartphone cameras with varying resolutions, and processed with multi spectral color features, rather than applying CNN models directly to the collected images. By simulating multispectral characteristics, this approach aims to improve pH detection accuracy from RGB images. In summary, this study explores the feasibility of smartphone-based pH prediction, examines the impact of camera resolution on prediction accuracy, and identifies suitable machine learning algorithms for this purpose.

MATERIALS AND METHODS

This research methodology was divided into four main steps: (1) Sample preparation and data collection, (2) Image Processing and features extraction (pre-processing data); (3) model development; and (4) evaluation. Flow diagram of proposed method for forecasting water quality parameter using image processing techniques and machine learning analysis shown in **Figure 1**.

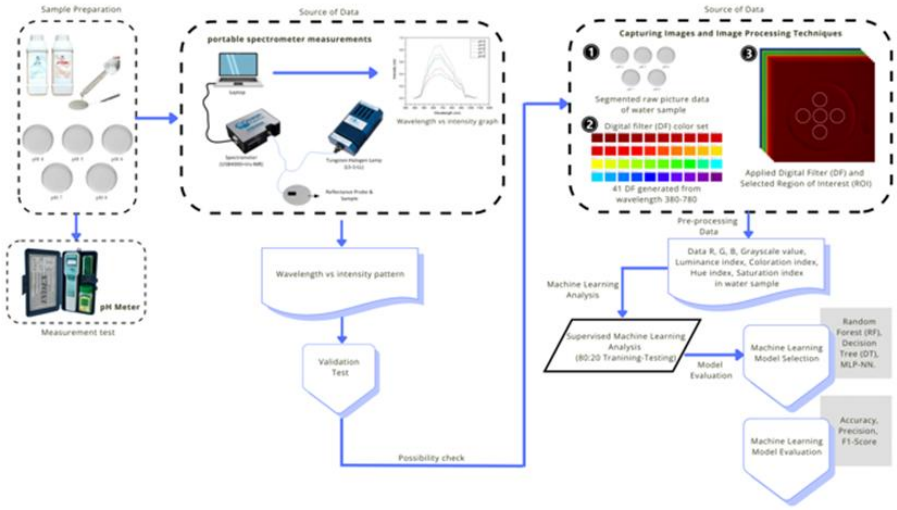


Figure 1. Research framework model classification development of forecasting water pH with generated multi spectral images and machine learning analysis

A. Study Area and Sample Preparation

Samples were collected in the Renewable Energy and Energy Conservation (REEC) laboratory, Chiang Mai University, Thailand. A total of 100 water samples were prepared by adjusting tap water pH using a 5% nitric acid (HNO_3) and 10% potassium hydroxide (KOH), covering a range of pH values 4 -8, in line with agricultural applications. Each sample was measured using pH meter Index (ID-1000 series).

B. Smartphone-based Digital Image and Multispectral Properties Simulation

The images of the samples were captured using two smartphone models: Samsung A22 with a 48 MP resolution camera and Samsung A03s with a 13 MP resolution camera. The samples were photographed from the top side to obtain a clear view of the water's surface, using white background to minimize noise and reflections. The lighting conditions were measured at 120 lux with digital light sensor LM-3000.

The images captured from the smartphone cameras have a resolution of 3088 x 3088 pixels. To simulate multispectral properties, a digital color filter was applied across a wavelength range of 380 to 780 nm, with increments of 10 nm, resulting in a total of 41 distinct digital filters. This approach allows each original image to be

converted into 41 separate images, each representing a different wavelength. The digital image processing was conducted using OpenCV with Python, facilitated through Google Collaboratory, enabling efficient manipulation and analysis of the multispectral data.

C. Image Processing and Features Extraction

Further analysis of image processing and feature extraction was conducted using ImageJ software. A total of 41 images were imported and analyzed using five regions of interest (ROI), each measuring 400 px x 400 px. From these ROIs, the R, G, B, and Grayscale values were extracted, resulting in a dataset where each sample produced 41 data points for each feature extraction. This process yielded a total of 1025 data points covering all target pH levels (4-8). In addition to the RGB and Grayscale values, further features such as luminance, hue, saturation, and coloration index were calculated using specific formulas derived from the extracted values. These features, summarized in Table 1, were crucial for enriching the dataset. All extracted features served as inputs for machine learning models, facilitating the prediction of water pH levels based on the multispectral imaging data.

Table 1. Feature extraction formula

Feature	Description	Source
Wavelength	Number of color filter	-
Red (R)	Extracted from ImageJ	-
Green (G)	Extracted from ImageJ	-
Blue (B)	Extracted from ImageJ	-
Grayscale	Formula: 0.299*R+0.587*G+0.114*B	ImageJ Software
Hue Index	Formula: (2*R – G – B)/(G-B)	[7]
Saturation Index	Formula: (R-B)/(R+B)	[7]
Coloration Index	Formula: (R-G)/(R+G)	[7]
Luminance Index	Formula: (0.2126 * R) + (0.7152 * G) + (0.0722 * B)	[8]

D. Machine Learning and Model Development

Machine learning methods, including support vector machines (SVM), random forests, decision trees, and neural networks, are widely used for their ability to recognize complex patterns and relationships within data. These algorithms learn from extensive training datasets to identify trends and make accurate predictions related to water quality. To ensure the accuracy and reliability of these models, they are tested with separate datasets. Techniques like k-fold cross-validation help assess model stability by evaluating performance across different data subsets. Additionally, using independent test datasets is crucial for verifying the model's ability to generalize to unseen data, a necessary step for real-world applications [9], [10].

For the analysis, supervised machine learning techniques were employed using three commonly used algorithms: Decision Tree (DT), Random Forest (RF), and Multi-Layer Perceptron Neural Network (MLP-NN). These algorithms were selected due to their effectiveness in handling complex datasets and their capability to model nonlinear relationships. To optimize the performance of these models, hyperparameter tuning was performed using RandomizedSearchCV, which efficiently explores a range of hyperparameter combinations to identify the best settings for each algorithm. The dataset was split into training and testing sets, with 80% allocated for training the models and 20% reserved for testing their predictive accuracy. This division ensures that the models are robust and can generalize well to unseen data. The performance of each algorithm was evaluated based on accuracy, precision, and F1-score, providing a comprehensive assessment of their effectiveness in predicting water pH levels from the

extracted features.

E. Model Evaluation

The evaluation of the machine learning models was conducted using several key performance metrics to assess their predictive accuracy and reliability. The primary metrics included accuracy, precision, recall, and F1-score, providing a comprehensive overview of each model's performance. Accuracy indicates the proportion of correctly predicted pH levels among the total predictions made. Precision measures the accuracy of the positive predictions, while recall assesses the model's ability to identify all relevant instances. The F1-score, which balances precision and recall, serves as a single metric to evaluate the model's overall effectiveness. Additionally, a confusion matrix was generated for each algorithm, allowing for a visual representation of the true positive, true negative, false positive, and false negative classifications. This matrix facilitates a deeper understanding of the models' strengths and weaknesses in predicting specific pH levels. By analyzing these metrics, we identified the most effective algorithm for predicting water pH levels and highlighted areas for further improvement in model performance. These metrics can be represented by the following equations:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{F1-Score} = \frac{2TP}{2TP+FN+FP} \quad (3)$$

where true positive (TP) is the positive data label that has been correctly predicted; A false positive (FP) is an incorrect prediction of a negative data label. A true negative (TN) is a correctly predicted negative data label, while a false negative (FN) is a positive data label that has been predicted incorrectly. Four widely recognized algorithms, namely ANN, SVM, decision tree (DT), and RF, are chosen for the comparative tests. The evaluation performance metrics, accuracy, precision, and F1-Score, are computed to assess the classification accuracies [11].

RESULT AND DISCUSSION

The findings from the analysis of spectral data and the performance of machine learning models used to predict water pH levels from camera smartphone-captured images were presented. The results are organized to first highlight the distinct spectral characteristics of pH-adjusted water samples, followed by a detailed evaluation of the model performance metrics, including accuracy, precision, and F1-score. Additionally, insights gained from the confusion matrix, feature importance analysis, and correlation matrix are discussed to provide a comprehensive understanding of how various factors contribute to the predictive capabilities of the developed model.

A. Spectral Data Analysis

The spectral characteristics of each pH-adjusted water sample were analyzed using an ocean brand portable spectrometer, resulting in distinct wavelength (nm) vs intensity (rel.) patterns. **Figure 2** illustrates the spectral data obtained for varying pH levels.

The initial dataset for spectral simulation included spectrometer data combine with the machine learning models demonstrated as with the simulated hyperspectral data [12]. The spectra of water with varying pH levels show distinct

reflectance patterns. The key point is that the acidity of the water significantly influences its spectral reflectance characteristics, with more acidic water showing reduced reflectance [13]

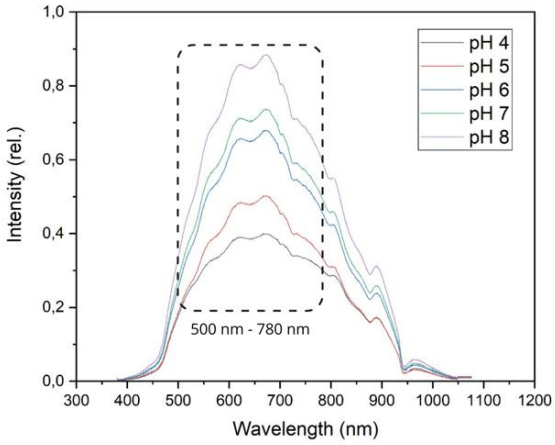


Figure 2. Spectral patterns of VIS-NIR for water samples at varying pH levels

The distinct wavelength vs. intensity patterns observed in the spectral data (Figure 3) suggest that pH levels can be effectively characterized through spectral analysis, especially in wavelength 500-780 nm. The variations in intensity and peak shifts support the hypothesis that spectral features are indicative of water quality changes. In line with A. Riaza [13], the acidity of the water significantly influences its spectral intensity characteristics, with more acidic water showing reduced intensity.

B. Feature Importance and Correlation Analysis

Feature importance analysis revealed the contribution of various extracted features in predicting pH levels. **Figure 3** displays the feature importance scores. Green and red values were identified as the most significant features, followed by wavelength, and then hue and coloration index.

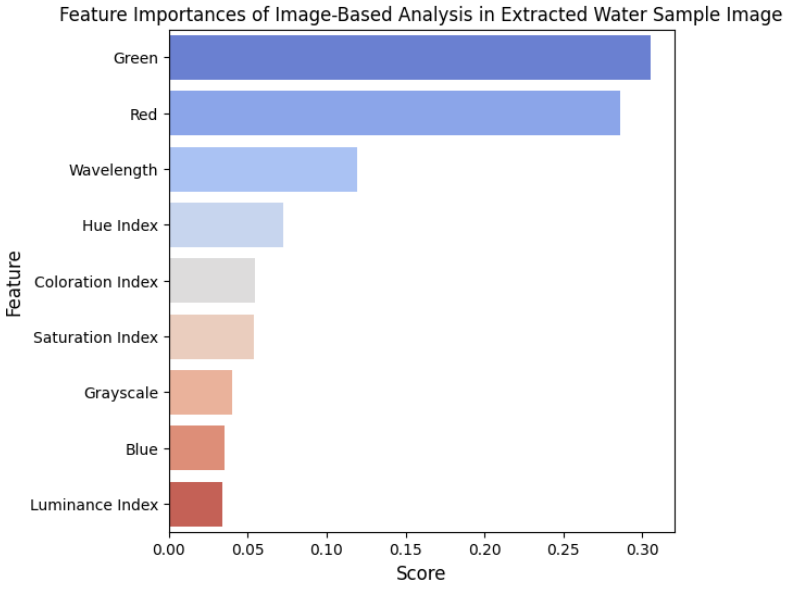


Figure 3. Feature importance score for pH prediction

The correlation matrix (Figure 4) revealed strong relationships between certain features, such as wavelength and coloration, suggesting that these features may be redundant in the model.

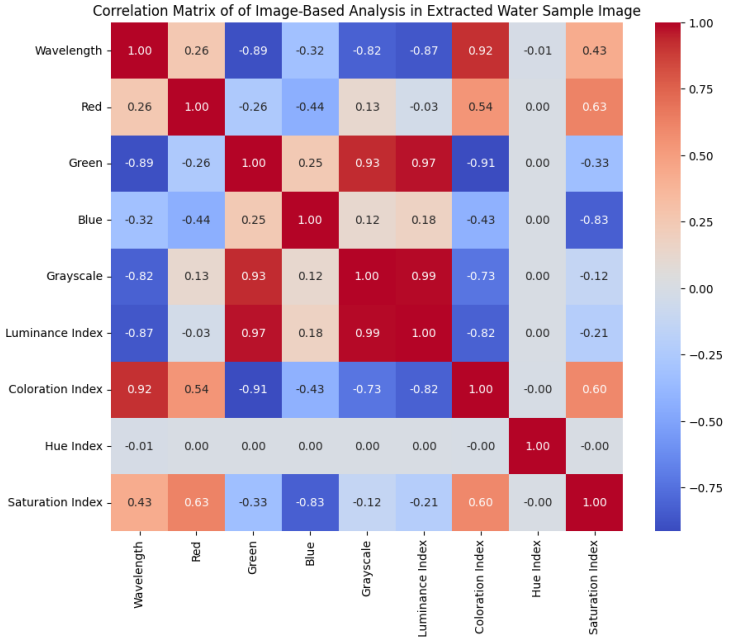


Figure 4. Correlation matrix for extracted features

The identification of RGB values as the most influential features align with previous research emphasizing the role of colorimetric data in pH prediction using digital image based. The strong correlation observed among extracted features suggest that simplifying the feature set may improve model efficiency without sacrificing accuracy. Future research could explore additional spectral bands or advanced feature extraction techniques to enhance the model's robustness.

C. Model Performance Metrics

The performance of the machine learning models was evaluated based on accuracy, precision, F1-score, and the confusion matrix. Table 2. summarizes these metrics for each model.

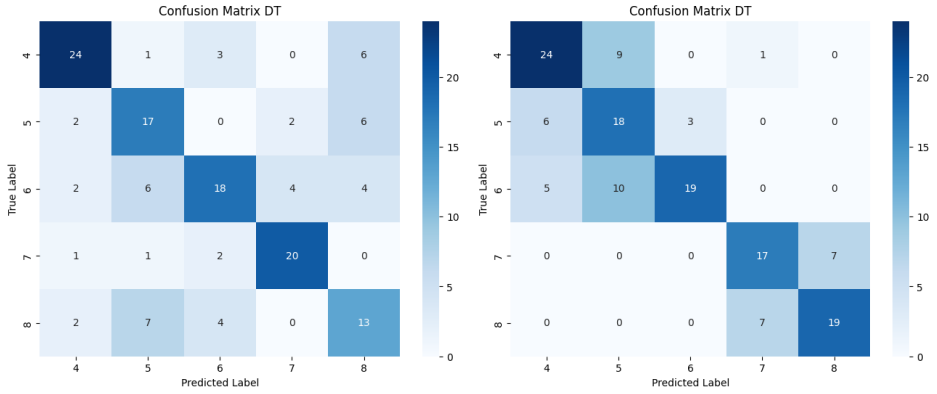
The Decision Tree model's superior accuracy (0.67) and precision (0.69) highlights its close effectiveness for this application. The comparable performance of the Random Forest model suggests that ensemble methods may be beneficial in similar studies, potentially enhancing predictive accuracy further. The F1-score results indicate a balanced model that performs well in both precision and recall, which is crucial for practical applications where both false positives and false negatives can have significant consequences.

Table 1. Performance metrics of machine learning models for pH prediction using 13 MP camera

Algorithm	13 MP Camera			48 MP Camera		
	Accuracy	Precision	F1-Score	Accuracy	Precision	F1-Score
Decision Tree (DT)	0.63	0.64	0.64	0.67	0.69	0.67
Random Forest (RF)	0.61	0.6	0.6	0.64	0.66	0.65
MLP-NN	0.62	0.64	0.62	0.6	0.6	0.6

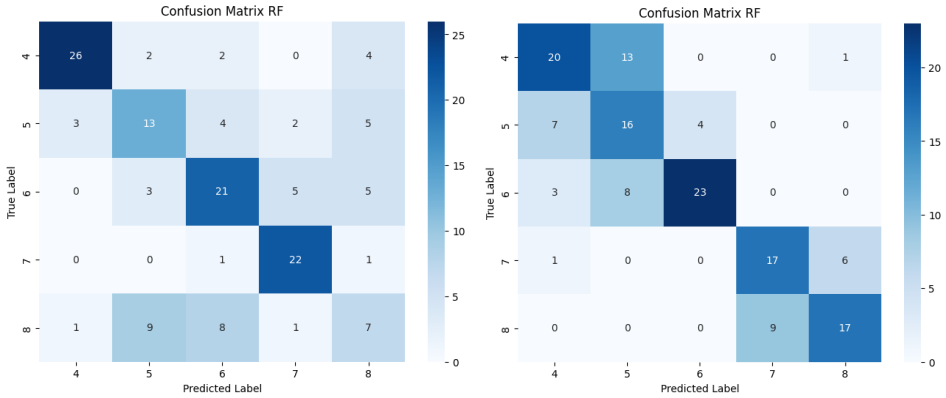
D. Model Performance Metrics

The confusion matrix for the Decision Tree (DT), Random Forest (RF) and MLP-NN model is presented in **Figure 5, Figure 6,** and **Figure 7**, illustrating the model's performance across different pH categories.



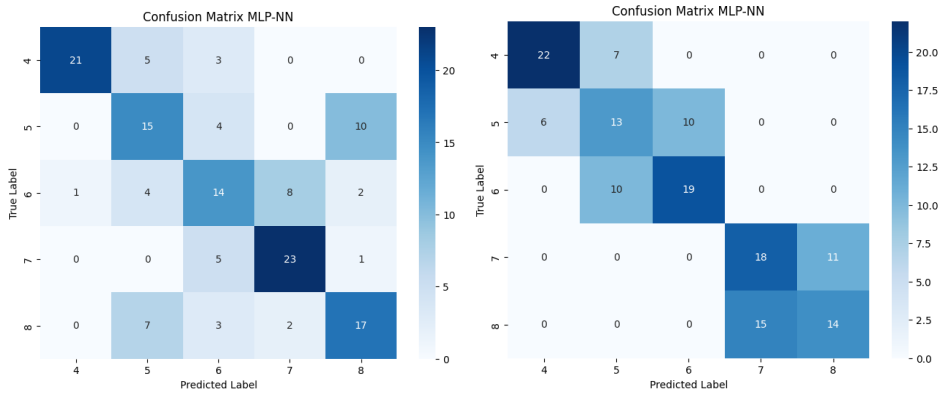
(a) (b)

Figure 5. Confusion matrix for the DT algorithm, using camera (a).13 MP and (b).48 MP.



(a) (b)

Figure 6. Confusion matrix for the RF algorithm, using camera (a).13 MP and (b).48 MP.



(a) (b)

Figure 7. Confusion matrix for the MLP-NN algorithm, using camera (a).13 MP and (b).48 MP.

The confusion matrix analysis points to the Decision Tree model's strengths and limitations, particularly in the pH transition areas. The challenges in classifying more large range of pH levels indicate a need for further refinement of the model, possibly through additional training data or advanced feature engineering. Nonetheless, factors such as the quality of the training data, the complexity of the water quality metrics, and the frequency of monitoring can influence this accuracy [14].

E. Limitation of This Study and Future Work

This study, while demonstrating the potential of smartphone-based imaging and machine learning for predicting water pH, has several limitations that suggest areas for improvement and future work. One key limitation is the relatively small dataset size, which may impact the model's generalizability and limit its ability to handle diverse real-world conditions. Expanding the dataset with additional pH levels and a wider range of water samples would enhance the robustness of the model. Additionally, the study utilized only two smartphone camera resolutions (13 MP and 48 MP), which limits the exploration of how different camera settings and smartphone models may affect prediction accuracy. Future work could involve a broader range of camera types and lighting conditions to test model performance under varied circumstances. Finally, the model's moderate accuracy suggests that integrating additional features, refining image-processing techniques, or exploring more complex machine learning algorithms may further improve prediction precision. Future research could focus on developing more sophisticated models and testing real-time pH prediction applications in diverse agricultural and environmental contexts.

CONCLUSION

This study demonstrates the feasibility of using smartphone cameras and machine learning models to predict water pH levels, presenting a practical approach that leverages digital imaging and accessible technology. By applying digital color filters across a simulated multispectral range, we extracted relevant features, such as RGB, grayscale, luminance, and saturation values, to train and test machine learning algorithms. Among the models tested, the Decision Tree algorithm showed the highest predictive accuracy, although improvements are needed for broader applicability. The results indicate that with further refinement, including larger datasets and enhanced feature selection, this method holds potential for field applications in agriculture and environmental monitoring. Ultimately, this study highlights a promising direction for smartphone-based pH prediction, laying the groundwork for future research to develop accurate, portable, and cost-effective water quality assessment tools.

Acknowledgements

The authors would like to express their sincere gratitude to the **Faculty of Engineering, Chiang Mai University**, for the financial support. The generous funding and resources have been instrumental in the successful completion of this study.

REFERENCES

[1] Kılıç, V., Alankus, G., Horzum, N., Mutlu, A. Y., Bayram, A., and Solmaz, M. E., 2018, Single-image-referenced colorimetric water quality detection using a smartphone. *ACS Omega*, 3(5), 5531-5536. doi: 10.1021/acsomega.8b00625.

[2] Liu, Y., Hu, X., Zhang, Q., and Zheng, M., 2017, Improving agricultural water use efficiency: A quantitative study of Zhangye City using the static CGE model with a CES water–land resources account. *Sustainability*, 9(2), 308.

doi: 10.3390/su9020308.

[3] Rostom, N. G., Shalaby, A. A., Issa, Y. M., and Afifi, A. A., 2017, Evaluation of Mariut Lake water quality using hyperspectral remote sensing and laboratory works. *The Egyptian Journal of Remote Sensing and Space Science*, 20, S39–S48. doi: 10.1016/j.ejrs.2016.11.002.

[4] Syahputra, W. N. H., Chaichana, C., Wongwilai, W., and Manggala, B., 2023, The use of neural network coupled with image processing for water quality assessment (Location: Hot Spring Mae-Khachan, Thailand). *International Energy Journal*, 23(1), 47–54.

[5] Wang, X., Zhang, F., and Ding, J., 2017, Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific Reports*, 7(1), 12858. doi: 10.1038/s41598-017-12853-y.

[6] Li, Y., Fu, Y., Lang, Z., and Cai, F., 2024, A high-frequency and real-time ground remote sensing system for obtaining water quality based on a micro hyper-spectrometer. *Sensors*, 24(6), 1833. doi: 10.3390/s24061833.

[7] Gozukara, G., Anagun, Y., Isik, S., Zhang, Y., and Hartemink, A. E., 2023, Predicting soil EC using spectroscopy and smartphone-based digital images. *Catena*, 231, 107319. doi: 10.1016/j.catena.2023.107319.

[8] Putra, B. T. W., Wirayuda, H. C., Syahputra, W. N. H., and Prastowo, E., 2022, Evaluating in-situ maize chlorophyll content using an external optical sensing system coupled with conventional statistics and deep neural networks. *Measurement*, 189, 110482. doi: 10.1016/j.measurement.2021.110482.

[9] Zhou, X., and Zhang, J., 2023, Advances in machine learning for water quality prediction and prospects in Erhai Lake. doi: 10.3233/ATDE230306.

[10] Wang, X., Li, Y., Qiao, Q., Tavares, A., and Liang, Y., 2023, Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy*, 25(8), 1186. doi: 10.3390/e25081186.

[11] Chen, J., Zhang, D., Yang, S., and Nanehkaran, Y. A., 2020, Intelligent monitoring method of water quality based on image processing and RVFL-GMDH model. *IET Image Processing*, 14(17), 4646–4656. doi: 10.1049/iet-ipr.2020.0254.

[12] Maier, P. M., and Keller, S., 2019, Application of different simulated spectral data and machine learning to estimate the chlorophyll A concentration of several inland waters. *Proceedings of the 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, IEEE, pp. 1–5. doi: 10.1109/WHISPERS.2019.8921073.

[13] Riaza, A., Buzzi, J., García-Meléndez, E., Carrère, V., Sarmiento, A., and Müller, A., 2015, Monitoring acidic water in a polluted river with hyperspectral remote sensing (HyMap). *Hydrological Sciences Journal*, 60(6), 1064–1077. doi: 10.1080/02626667.2014.899704.

[14] Jayaraman, P., Nagarajan, K. K., Partheeban, P., and Krishnamurthy, V., 2024, Critical review on water quality analysis using IoT and machine learning models. *International Journal of Information Management Data Insights*, 4(1), 100210. doi: 10.1016/j.jjimei.2023.100210.