Prediction system for pH measurement on Brassica oleraceae (Red Cabbage) using machine learning regression

Dian Wulan Hastuti, Marlina Harahap, Faizal Adila Ferdiansyah, Adhi Harmoko Saputro, and Cuk Imawan*

Department of Physics, Faculty of Science and Mathematics, Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

* cuk.imawan@sci.ui.ac.id

Abstract. pH is an important unit to represent the chemical condition of a liquid, solid substance, food nutrition, and microbial activity as well. pH value is also commonly used to detect the behaviour of chemical substances. Measurement of pH value can be shown by color changes of the substance based on the acidity condition from the measured environment. In this research, the colorimetric based on machine learning and pH value detected from the color information for point-of-care applications. For this investigation, we used the pH buffer solution and natural dyes derived from Brassica oleraceae (Red Cabbage) that shows colorimetric response gradually shifted from red to green along with the increasing of pH from 2.00 (acid) to 11.00 (alkaline). In this paper, we propose a method for predicting pH value based on Artificial Neural Network Regression (ANNR) and K-Nearest Neighbour Regression (KNNR) with RGB, HSV and LAB color space. As a result, the performance (99.83% ± 0.11) of this method could estimate the pH reasonably well for point-of-care application.

1. Introduction

The various important of pH is to represent the chemical condition of a liquid and solid substance [1], bioanalytical activity [2]–[5], health monitoring [5]–[7] and environmental concern [8]–[11]. To overcome the intrinsic image sensor properties, a calibration technique is commonly applied with a real accurate pH buffer or a high-quality verified pH measurement device[11], [12].

Color change of a substance can indicate the pH value and becomes essential for classification element [12]. Color has also been closely associated with various quality factors such as food safety, water quality or human fluids to detect health quality.

The study of natural dyes is an active and broad area because it supports the interest in substituting synthetic dyes that have toxic effects on the human body. Anthocyanin is the most important pigment of the vascular plants due to the harmless and easy incorporation in aqueous media, which makes it interesting as a natural water-soluble colorant [13].

The color adaptation has been used to inspect the visual perception related to the identification and classification of the problems. Wu and Sun (2013) have classified three spaces of color; hardware-orientated spaces (RGB, YIQ, and CMYK), human-orientated spaces (HSI, HSL, HSV, and HSB), and instrumental spaces (XYZ, L*a*b, and L*u*v) [13], [14], [17].

By introducing the machine learning algorithm, several pattern recognition techniques in image analysis can be used. The two types of machine learning training are the supervised and unsupervised
method. The supervised methods like the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), and decision tree (DT) are the most commonly used modelling utilization for classification and regression problems [14]–[18]. Meanwhile, the unsupervised methods, learning methods to build a machine learning model, where training data is not accompanied by targets or labels like the K-Means Clustering, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for clustering and dimensionality reduction utilization.

2. Equipment and Materials

2.1. Red Cabbage (Brassica oleraceae) Extraction
Natural dyes derived from Red Cabbage has been prepared to investigate the existence of anthocyanin (Figure 1). The solution was prepared at a concentration 0.2 gr/ml on the aqueous solution with various pH values by using 1 M (molar) solution of Hydrochloric acid (HCl) and Sodium Hydroxide (NaOH) then obtained pH value from 2.00 (acid) to 11.00 (alkaline) [19]–[21].

![Figure 1. Characterization of Anthocyanin Dyes on Brassica oleraceae.](image)

2.2. Equipment’s and Computer Vision
Below is the hardware and software used in the experiment:
- Photo box studio. The size of the box is 470x340x330 mm.
- Light source. Two long fluorescent tubes with illumination ≈ 3750 LUX placed at the top and bottom of the photo studio box.
- Digital camera. Using Canon™ EOS 750D with resolution 6000x4000 pixels. The digital camera was located ≈ 24 cm in front of the sample inside the photo box studio.
- Computer. Using a personal computer (Intel(R) Core™ i7-2600 CPU @ 3.40 GHz and 4.0 GB RAM)
- Data Processing. For pre-processing, data extraction, feature selection and modelling of the image are implemented using Python application tools.

3. Method
The experimental methodology was based on Castro et al [22]. In general, input data of RGB image captured had first extracted to obtain feature vectors by each vector space and classification models. Then, the data is separated into training and testing data by using cross-validation method to improve the prediction using the K-Fold strategy. The data extraction method was implemented to find the best metric performance between two regression models, K-Nearest Neighbor (KNNₖ) and Artificial Neural Network (ANNₖ).

3.1. Pre-processing
In this step, 8 captured images which consists of RGB Spectrum has an optical resolution of 6000x4000 pixels which are separately cropped by each tube of the pH value from 2.00 (acid) to 11.00 (alkaline) and obtained 10 Region of Interest (ROI). Each of these ROIs has 450x600 pixel resolution for each pH value (totally generated of 8x10 images was composed of the three-color channels, Red, Green and Blue). The final step was cropping every 8x10 ROI by dividing into 5 ROI then becomes 8x10x5 dataset which has resolution 90x600 pixel of each ROI.

3.2. Data extraction
Input data of Red, Green, Blue (RGB) channel image then converted to another color space. In this research, each pixel of images was converted into the HSV and LAB color space using the OpenCV Python library (RGB2HSV and RGB2LAB).

3.3. Feature Selection
One of the popular unsupervised machine learning is K-Means Clustering. K-Means Clustering is an unsupervised learning algorithm that allows to cluster similar data points within entire data to a group. The data is grouped into K clusters by measuring distance or similarity to a centroid that is randomly generated. Then the centroids are computed several times as the mean of the data points to the respective cluster, until it reaches the determined criteria. It has been used to find dominant values in the selected area of 450x600 pixels. In this step, the training and testing data with 5-fold cross-validation strategy using color features of 8x10x5 were used to obtain 5 dominant color from each color space.

3.4. Modelling
Two model techniques were adopted, namely, the K-Nearest Neighbor Regression (KNNR) and Artificial Neural Network Regression (ANNR). Each combination of color space (RGB, HSV and LAB) and regressor were compared by the performance measure.

The K-Nearest Neighbor (KNNR) is a simple supervised algorithm of machine learning that has the purpose to resolve the problems off classification and regression. The KNN algorithm works by finding for related things so that they are at close range that can be referred to as "Neighbors".

The artificial neural network is used to detect pH measurement based on neural network representation. ANNR is the Feedforward Neural Networks (FNN), which is the model of a neural network where the connections between neurons do not form a cycle. The artificial neural network consists of three processing layers namely, input layer, hidden layer and output layer. The input layer defines the data input given to the neural network processes. the hidden layer or called middle layer,
there is a non-linear activation function. In this research, ReLU is used as an activation function. The last layer is the output layer which represents the result, the predicted of pH measurement.

4. Result and Discussion

This section present experiment of prediction system for pH measurement of anthocyanin dyes from *Brassica oleraceae* (Red Cabbage) at ten measured pH values from pH from 2.00 (acid) to 11.00 (alkaline).

**Table 1.** The average performance of each model and color space

| Regressor | RGB Color space | Cross-validation average value based on dominant color clustering |  |  |  |
|-----------|-----------------|---------------------------------------------------------------|---|---|---|
|           |                 | Training                                                      | Testing | R-Squared | RMSE | Correlation Coefficient | R-Squared | RMSE | Correlation Coefficient |
| KNN_R     | 0.99954         | 0.06053                                                      | 0.99977 | 0.99831 | 0.11476 | 0.99918 |
| ANN_R     | 0.95141         | 0.62908                                                      | 0.97668 | 0.94088 | 0.69268 | 0.97266 |

| Regressor | HSV Color space | Cross-validation average value based on dominant color clustering |  |  |  |
|-----------|-----------------|---------------------------------------------------------------|---|---|---|
|           |                 | Training                                                      | Testing | R-Squared | RMSE | Correlation Coefficient | R-Squared | RMSE | Correlation Coefficient |
| KNN_R     | 0.99934         | 0.07370                                                      | 0.99967 | 0.99745 | 0.13753 | 0.99877 |
| ANN_R     | 0.71739         | 1.52331                                                      | 0.85912 | 0.66026 | 1.62219 | 0.83261 |

| Regressor | LAB Color space | Cross-validation average value based on dominant color clustering |  |  |  |
|-----------|-----------------|---------------------------------------------------------------|---|---|---|
|           |                 | Training                                                      | Testing | R-Squared | RMSE | Correlation Coefficient | R-Squared | RMSE | Correlation Coefficient |
| KNN_R     | 0.99945         | 0.06618                                                      | 0.99972 | 0.99715 | 0.14919 | 0.99860 |
| ANN_R     | 0.96036         | 0.56617                                                      | 0.98048 | 0.94325 | 0.67484 | 0.97292 |

Table 1 summarizes the performance of each model the K-Nearest Neighbor (KNN_R) and Artificial Neural Network (ANN_R) and each color space RGB, HSV and LAB in term of R-squared, RMSE and correlation coefficient. Both KNN_R and ANN_R has already been successfully used to validate pH according to their value 2.00 to 11.00 using RGB, HSV and LAB color space.

The bad performance resulted from HSV color space using the ANN_R modelling yielded the R-squared, RMSE and correlation coefficient value respectively using cross-validation on data testing was 66.03%, 1.62, and 0.83 based on dominant color clustering. The KNN_R have achieved results with best performance levels on RGB color space on the R-squared, RMSE and correlation coefficient value respectively was 99.83%, 0.11 and 0.99.
5. Conclusion
The purpose of this research was to develop a system for the prediction of pH value from an acid (2.00) to alkaline (11.00) using Anthocyanin natural dyes which extracted from *Brassica oleracea* (Red Cabbage). Six classification models were developed by utilizing two machine learning, the K-Nearest Neighbor Regression (KNN<sub>R</sub>) and Artificial Neural Network Regression (ANN<sub>R</sub>) with three color spaces (RGB, HSV and LAB) as its feature.

K-Means Neighbor has been used to find dominant values on each ROI sample that employed all-of-three color spaces. Meanwhile, the models based on the KNN<sub>R</sub> and ANN<sub>R</sub> yielded adequate results with machine learning techniques employed. The KNN<sub>R</sub> achieved results with best the performance levels of R2 and RMSE, a coefficient correlation on RGB color space value respectively was 99.83%, 0.11 and 0.99.

References
[1] S. M. Ghoreishian *et al.*, “γ-Radiolysis as a highly efficient green approach to the synthesis of metal nanoclusters: A review of mechanisms and applications,” *Chem. Eng. J.*, vol. 360, pp. 1390–1406, 2019.
[2] S. Álvaro *et al.*, “A personalized food allergen testing platform on a cellphone,” *Lab Chip*, vol. 15, no. 207890, pp. 4491–4498, 2013.
[3] M. Citartan, Y. Chen, T.-H. Tang, T. Lakshmipriya, and S. C. B. Gopinath, “Bacterial detection: From microscope to smartphone,” *Biosens. Bioelectron.*, vol. 60, pp. 332–342, 2014.
[4] E. M. Schneider, S. K. Vashist, J. H. T. Luong, R. Zengerle, T. van Oordt, and F. von Stetten, “A smartphone-based colorimetric reader for bioanalytical applications using the screen-based bottom illumination provided by gadgets,” *Biosens. Bioelectron.*, vol. 67, pp. 248–255, 2014.
[5] Z. D. and L. Q., “Biosensors and bioelectronics on smartphone for portable biochemical detection,” *Biosens. Bioelectron.*, vol. 75, pp. 273–284, 2016.
[6] M. Y. Jia, Q. S. Wu, H. Li, Y. Zhang, Y. F. Guan, and L. Feng, “The calibration of cellphone camera-based colorimetric sensor array and its application in the determination of glucose in urine,” *Biosens. Bioelectron.*, vol. 74, pp. 1029–1037, 2015.
[7] S. S. Shivanni, V. Vishnupriya, and R. Gayathri, “Analysis of lipid profile in saliva of diabetic patients - A pilot study,” *J. Pharm. Sci. Res.*, vol. 8, no. 6, pp. 477–479, 2016.
[8] J. P. Devadhasan, D. Kim, D. Y. Lee, and S. Kim, “Smartphone coupled handheld array reader for real-time toxic gas detection,” *Anal. Chim. Acta*, vol. 984, pp. 168–176, 2017.
[9] X. Meng, C. W. Schultz, C. Cui, X. Li, and H. Z. Yu, “On-site chip-based colorimetric quantitation of organophosphorus pesticides using an office scanner,” *Sensors Actuators, B Chem.*, vol. 215, pp. 577–583, 2015.
[10] M. L. Firdaus et al., “Smartphone Coupled with a Paper-Based Colorimetric Device for Sensitive and Portable Mercury Ion Sensing,” 2019.

[11] S. Schaefer, “Colorimetric water quality sensing with mobile smart phones,” no. April, p. 88, 2014.

[12] S. D. Kim, Y. Koo, and Y. Yun, “A smartphone-based automatic measurement method for colorimetric pH detection using a color adaptation algorithm,” Sensors (Switzerland), vol. 17, no. 7, 2017.

[13] E. Alexandra Pazmio-Durán, M. M. Giusti, R. E. Wrolstad, and M. B. A. Glória, “Anthocyanins from banana bracts (Musa X paradisiaca) as potential food colorants,” Food Chem., vol. 73, no. 3, pp. 327–332, 2001.

[14] W. Castro, J. Oblitas, M. De-La-Torre, C. Cotrina, K. Bazan, and H. Avila-George, “Using Machine Learning Techniques and Different Color Spaces for the Classification of Cape Gooseberry (Physalis peruviana L.) Fruits According to Ripeness Level,” IEEE Access, vol. 7, no. 2006, pp. 27389–27400, 2019.

[15] P. T. Wang, J. J. Chou, and C. W. Tseng, “Colorimetric characterization of color image sensors based on convolutional neural network modeling,” Sensors Mater., vol. 31, no. 5, pp. 1513–1522, 2019.

[16] Z. M. Khaing, Y. Naung, and P. H. Htut, “Development of control system for fruit classification based on convolutional neural network,” Proc. 2018 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. ElConRus 2018, vol. 2018-Janua, pp. 1805–1807, 2018.

[17] D. Wu and D. W. Sun, “Colour measurements by computer vision for food quality control - A review,” Trends Food Sci. Technol., vol. 29, no. 1, pp. 5–20, 2013.

[18] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with exemplar convolutional neural networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 9, pp. 1734–1747, 2016.

[19] S. Pourjavaheh, H. Almasi, S. Meshkini, S. Pirsa, and E. Parandi, “Development of a colorimetric pH indicator based on bacterial cellulose nanofibers and red cabbage (Brassica oleracea) extract,” Carbohydr. Polym., vol. 156, pp. 193–201, 2017.

[20] N. G. Akram, W. Afridi Bhattu, and I. Nawaz Sharif, “A study on the response of natural dye to Gamma radiation as a dosimeter,” 2016.

[21] I. N. Handayani and C. Imawan, “Liquid Radiocromic from Roselle Dye Extract for Gamma-ray Dosimetry Applications at Medium Dose Levels,” in ISSIMM 2018 - 3rd International Seminar on Sensors, Instrumentation, Measurement and Metrology. Proceeding, 2018, pp. 115–118.

[22] W. Castro, J. Oblitas, J. Maicelo, and H. Avila-George, “Evaluation of expert systems techniques for classifying different stages of coffee rust infection in hyperspectral images,” Int. J. Comput. Intell. Syst., vol. 11, no. 1, pp. 86–100, 2018.