Proceedings Paper:
Wang, Y., Evans, C., Zhang, L. orcid.org/0000-0002-4535-3200 et al. (1 more author) (2024) Machine Learning and Colorimetric Method Based pH Detecting System. In: ICRCV 2024 Conference Proceedings. 2024 6th International Conference on Robotics and Computer Vision, 20-22 Sep 2024, Wuxi, China. IEEE , pp. 353-358. ISBN 979-8-3315-2743-3

https://doi.org/10.1109/ICRCV62709.2024.10758594

Machine Learning and Colorimetric Method Based pH Detecting System

Yuzhe Wang, Craig A. Evans, Li Zhang, Farah Al-Sallami
School of Electronic and Electrical Engineering
University of Leeds
Leeds, United Kingdom
{el22yw, C.A.Evan, L.X.Zhang, and F.Al-Sallami}@leeds.ac.uk

***Abstract*—This study investigates machine learning-based methods to measure the pH level accurately while using inexpensive equipment. The colorimetric method is employed to determine the pH level of a solution, which is achieved through a pH indicator paper-based embedded test system. The system incorporates a machine vision module to identify the colour of the pH indicator paper and a machine learning algorithm to quantify the pH value. A comparison between regression and classification machine learning algorithms was conducted. The experimental results revealed that, despite the regression model exhibiting smaller pH intervals than the classification model, the classification model is more stable and estimates more accurate values.**

***Index Terms*—machine learning, machine vision, pH measurement**

I. Introduction

pH refers to the potential hydrogen, which is the critical indicator for measuring the acidity or alkalinity in water-soluble substances. Accurate measurement of the pH-level is essential in various fields, such as agricultural production, food processing, and medical manufacturing.

A range of test equipment are used for pH-value detection. The glass electrode pH meter and the semiconductor pH sensor are both highly accurate [1]. However, the glass electrode meter has poor mechanical strength, and the semiconductor sensor has a relatively short lifespan. Furthermore, both the glass electrode pH meter and the semiconductor pH sensor are considerably expensive. The pH indicator is inexpensive, but the estimated value is highly dependent on the subjective judgement of the user. The colour grade of the pH indicator is not uniformly distributed. The data from the tests [2], [3], indicates that significant shifts in the pH of the test solution result in only minor changes in the absorption wavelength of the pH indicator. Consequently, only minor changes in the colour of the pH indicator are observed when there are significant pH differences in the test solution. Moreover, environmental factors such as the colour of the test solution or lighting conditions can influence the accuracy of pH test results obtained using a pH indicator. It is, therefore, essential to consider these factors to avoid errors in pH measurement.

Machine learning-based techniques were proposed in the literature to determine the pH value through identifying the colour of the pH indicator. In [4], [5], [6], regression machine learning algorithm was demonstrated to be effective in this context. In [7], [8], [9], a classification algorithm was employed with precise pH indicator paper, which contains multiple targets in a single sample, resulting in more significant colour differentiation for different pH values. In [10], the interpolation technology is employed to generate the dataset on the basis of the theoretical pH-colour relationship.

Accordingly, this study employs a series of classification and regression machine learning algorithms, in conjunction with the machine vision function, to ascertain the pH value from the pH indicator. To this end, a 3D-printed test machine with an embedded system has been constructed, offering an inexpensive, high-accuracy pH detection solution.

II. Machine Learning Model

This section provides a comparison between the performance of classification and regression algorithms in performing accurate colorimetric tests.

A. Machine Learning Models

In this study, a number of commonly used machine learning models were selected for both classification and regression, including Decision Trees (DT), K-Nearest Neighbour (KNN) for classification, linear regression (LR), Regression Trees (RT) for regression, and Support Vector Machines (SVM) and neural Networks (NN). In addition to the Gaussian Radial Basis Function (RBF) kernel in the SVM that is utilized in classification algorithm.

B. Colour Space Conversion

The colour space conversion, as a pre-processing step for a machine learning program, exerts a profound influence on the accuracy and performance of the machine learning model, particularly when the machine learning model is deployed on an embedded system. This study examines four colour spaces: Red-Green-Blue (RGB), Hue-Saturation-Value (HSV), International Commission on Illumination L*a*b* (CIELAB), and International Commission on Illumination XYZ (CIEXYZ). The study assesses the influence of each colour space on machine learning models.

C. Training Datasets

Classification algorithm is suggested when the colour data in the dataset used to train the machine learning model is discrete.

However, as the pH interval decreases, the distance between data points also decreases, resulting in a linear distribution of data within the dataset. Therefore, before evaluating the performance of the machine learning model using different algorithms, selecting a suitable dataset to train the machine learning model and investigating the effect of the colour space on the dataset are essential requirements. The study utilises three independent datasets from different sources and investigates four different colour spaces. The colour data in the dataset will be converted to different colour spaces using mathematical methods and then evaluated in different colour spaces.

TABLE I: Comparison of Three Datasets

	Dataset-1 [11]	***Dataset-2 [4]***	***Dataset-3 [10]***
Source	Online	Test result	Theoretical value
Data samples	653	128	133
pH range	0:14	0.7:14	0:14
pH interval	±1	±0.1	±1

Table 1 shows that Dataset-1, obtained from Kaggle, contains 653 test samples [11]. The dataset covers the entire pH range from 0 to 14 with an interval of 1. However, the origin of the data points is unclear, and there is no information to indicate whether the data was obtained through experimentation or generated by software. Dataset-2, obtained through experimentation, has a smaller pH interval than the others [4]. However, only a small portion of the test data points are available, and test data does not cover the entire pH range, starting from 0.7 to 14 instead of 0 to 14. The final dataset used in this study, Dataset-3, is based on the relationship between the ideal colour and the pH value, as indicated in [10]. The Dataset-3 is constructed using the interpolation process in MATLAB, resulting in 133 data points with a pH interval of 1 for classification algorithm and 141 data points with a 0.1 pH interval for regression algorithm.

Figure 1 depicts the efficacy of the LR algorithm at varying pH intervals utilised for training the regression model. The performance of the regression machine learning model is evaluated using a series of metrics, including the Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), which indicate the error and fit effect for the training model [4]. As the pH interval decreases, the performance of the regression algorithm improves, as evidenced by a reduction in RMSE, MSE, and MAE, and an increase in R^2. The performance of the linear regression machine learning model demonstrates a significant improvement as the pH interval decreases from 1 to 0.1, while after the 0.1 pH interval, the performance showed only a limited increase as the pH interval decreased further. However, as the pH interval decreases, which means that the amount of data in the dataset increases, the cost of training the machine learning model will increase. Therefore, considering the cost-performance balance, this study chose the 0.1 pH interval data as the dataset to train the regression machine learning model.

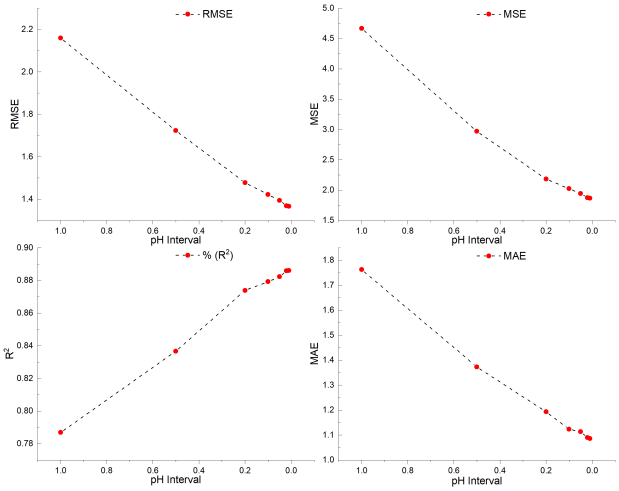

Fig. 1: Relation between linear regression performance and interval of pH.

D. Model Training Result

The machine learning model proposed in this research is trained in Figure 2 illustrates the accuracy of various trained classification machine learning models for all datasets. The parallel plots demonstrate that the results of Dataset-2 and Dataset-3 exhibited a notable enhancement in performance compared to Dataset-1. The HSV colour space for the trained machine learning algorithms exhibits the most stable and high-performing results. For the HSV colour space, Dataset-2 achieved the highest accuracy in the SVM algorithm (0.93), and Dataset-3 achieved the highest accuracy in the DT algorithm (0.92). However, the highest accuracy of the Dataset-2 and Dataset-3 training results appear at the RGB colour space for different algorithms achieved by the SVM algorithm (0.95) for Dataset-2 and the KNN algorithm (0.94) for Dataset-3.

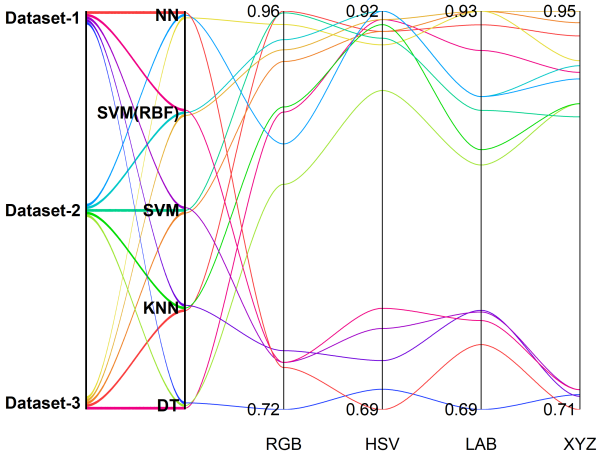

Fig. 2: Classification algorithm training result.

The regression machine learning model training result is illustrated in Figure 3. Dataset-3 is used with the interpolation process to create the 0.1 pH interval dataset. Figure 3 illus-

trates that the RT algorithm demonstrated the most optimal performance, with an RMSE of 0.32 and an R^2 of 0.99. The LR algorithm also demonstrated satisfactory performance, with an RMSE of 1.42 and an R^2 of 0.88. Notwithstanding the satisfactory performance of the NN, and SVM regression algorithms, a significant discrepancy remains between their outcomes and those of the optimal results.

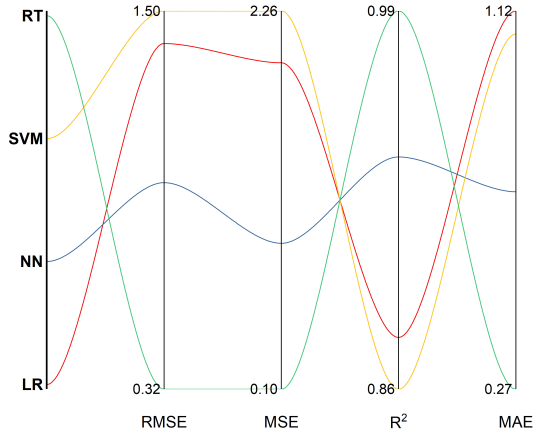

Fig. 3: Regression algorithm training result.

III. Machine vision function

A. Machine vision algorithm

This study introduces a machine vision function for the detection of the colour of a target pH indicator, which comprises several components. In the image blurring process, a median filter is applied to blur the image, smoothing the data and removing noise. Subsequently, OTSU method [12] is employed in the image binarization process to identify the Region of Interest (ROI) of the input image, which corresponds to the coordinates of the pH indicator paper in the image. The coordinates of the pH indicator paper are employed in the blob analysis process to ascertain the average colour value for the ROI. The final process for the machine vision algorithm is to convert the detected colour value to the RGB format.

B. Image Binarization Algorithm

The binarization algorithm is utilised to identify the coordinates of the ROI within the input image, thereby enabling the application of ROI coordinates statistics to the colour within the ROI. In order to implement the binarization process, the OTSU method is employed to determine the optimal threshold, which involves the separation of pixels in an image into two classes: background and foreground [12]. The maximised variance between these classes is then identified as the threshold, which effectively separates the objects of interest from the background.

Figures 4a) and 4b) demonstrate the effectiveness of the OTSU method in delineating clear foreground and background regions. However, the binarized images exhibit errors resulting from non-uniform illumination intensity. Figures 4c) and 4d) illustrate that the algorithm is less effective in images where

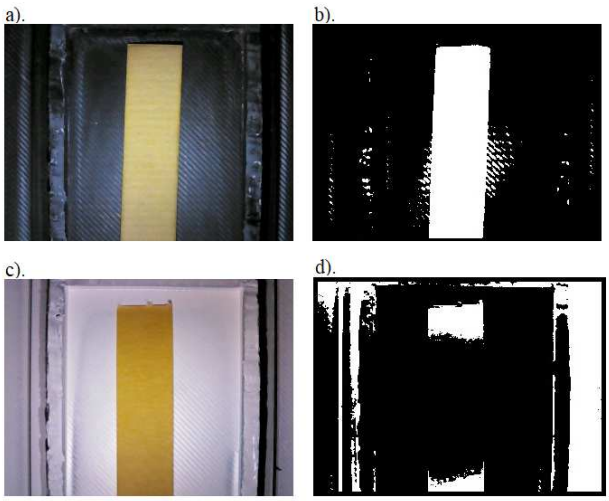

Fig. 4: OTSU method testing result a) captured pH indicator paper in black background, b) binarization results for black background, c) captured pH indicator in white background, and d) binarization results for white background.

the separation between the foreground and background regions is unclear. This is due to the similarity in colour between the light-coloured target and the white background, which makes it difficult for the algorithm to distinguish between the two regions. Similarly, the same issue arises when the dark-coloured target and the black background are used.

IV. Test Machine Design

A test machine has been designed and constructed as part of this study with the objective of verifying the machine learning algorithms and machine vision process. The test machine provides a controllable test environment for the machine vision module, thus eliminating the interference from ambient lighting. The chassis design is inspired by [13]. Furthermore, it serves as the reference lighting for the colour detection process. Figures 5 illustrate the design and build of the test machine. As demonstrated in Figure 5a), the components *a*, *b*, and *d* of the test machine construct the chamber to block ambient light and provide an optical path for the machine vision module. Component *c* is the test tray for containing the testing pH indicator paper. Components *e*, *f*, *g* focus on providing internal Light-emitting Diode (LED) lighting source and making the internal lighting suitable for the machine vision module. Finally, components *h*, *i*, *j* are used to support system operation.

However, the structure of the chamber of the test machine and the lighting source result in a non-uniform distribution of light on the test target, which introduces an error in the machine vision colour detecting process due to non-uniform illumination. Moreover, the impact of coloured solutions and illumination on the colour also affects the detection process [13]. Consequently, several correction methods are introduced in this study, as detailed in following sections.

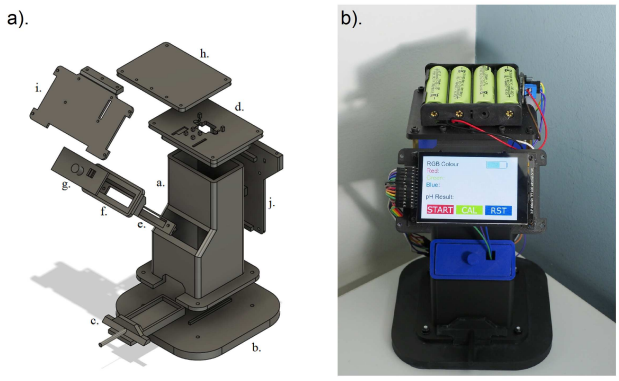

Fig. 5: Design and build of the test machine, a) test machine design, b) test machine build.

A. Lighting Correction

This study proposes a method for redistributing the light flow and resolving the illumination problem. The TacePro simulation was utilized to design a planar reflective diffuser plate to be attached to the front of the LED lighting source to provide the most uniform gradient distribution of illumination [14]. The results of the simulation, depicted in Figure 6, illustrate the luminance map for the pH indicator paper detecting area with and without the light diffuser component and with different angles of the light diffuser. Figure 6a) depicts the simulation results for the test machine in the absence of the light diffuser. Due to the configuration of the test apparatus, the region of interest, as illustrated in Figure 6a), which is visible in the machine vision module detection area, is smaller than the entire tray area.

In the absence of the light diffuser, the luminance distribution on the tray area exhibits a non-uniform distribution, with a significant illumination gradient difference. The maximum irradiance difference between luminance levels within the ROI is $112W/m^2$. Figure b) to f) illustrates the planar reflective diffuser plate integrated into the test apparatus, with the angle between the diffuser and the light source increasing from 30 to 135 degrees. As illustrated in the Figures, the diffuser block structure effectively blocks direct light and improves the diffusion of light to the test area, thereby improving uniform light distribution and reducing the gradient of the luminous flux. However, as the degree of the diffuser further increases, as illustrated in Figures 6d), e) and f), the proportion of direct light in the test area increases, resulting in an increase in irradiance and a rise in the illumination gradient difference.

In particular, the 30-degree light diffuser achieves optimal uniformity of light distribution on the test area with a mere $1.85W/m^2$ irradiance difference. The diffuser plate enhances the efficacy of machine vision algorithms by mitigating the non-uniform light gradient. As demonstrated in Figure 7, the 30-degree light diffuser limits the optical path for direct light within the test machine and eliminates direct light (red rays in the Figure) in the test area. The illumination in the test area is derived from the diffusion light (blue rays in the Figure), thereby creating an almost uniform light distribution in the test area. In this study, a 30-degree planar reflective diffuser light diffuser prototype was constructed utilising black paper, which demonstrated optimal performance in terms of light diffusion. However, the material of the light diffuser surface also demonstrates the impact of the real situation, as different materials have different reflectivity, which in turn affects the optical path inside the test machine.

B. Background Correction

As illustrated in Figure 4, the binarization algorithm is susceptible to inaccuracy when distinguishing the target colour from a similar background. This limitation can be addressed by employing alternative algorithms. Consequently, a correction method is employed in this study utilising a grey background. In order to address the aforementioned disadvantage, a series of grey backgrounds were tested and evaluated in this study. The colour gradient of the grey background ranged from total white to total black. The optimal RGB value for a grey background is presented below, which was obtained using the Datacolor ColorReader®.

$$Colour_{(R,G,B)} = (117, 113, 121) \tag{1}$$

The binarization algorithm after using the grey background, has the capacity to distinguish between the target and the background with greater clarity, thereby enhancing the accuracy of the machine vision model.

C. Colour Calibration

A colour calibration is introduced into the system to calibrate the colour detected by the machine vision module and provide accurate feedback [15].

$$C_{cal} = \frac{256}{C_w - C_b}(C_{mea} - C_b) \tag{2}$$

To calibrate the RGB colour, the calibration tape is employed, which comprises both a pure white and a pure black section, as illustrated in Figure 8. The reference white colour ((C_w) and black colour (C_b) are detected inside the test machine under the same illumination conditions. The uncalibrated colour (C_{mea}) is then detected separately by the machine vision module, which outputs the RGB format and comprises three colour channels, namely red, green and blue. Hence, the calibrated colour (C_{cal}) can be derived by calculating the individual channels of the detected colour and the reference black and white colour. This calibrated colour avoids the interference of the illumination and coloured solution, thus ensuring its suitability for use in machine learning algorithms to obtain the accurate pH value.

V. Test Result and Discussion

In this study, three test samples employed in this study to assess the pH value as demonstrate in Figure 9. The pH buffer solution is utilized as it can maintain high accuracy in pH value across diverse conditions. The three pH buffer solutions

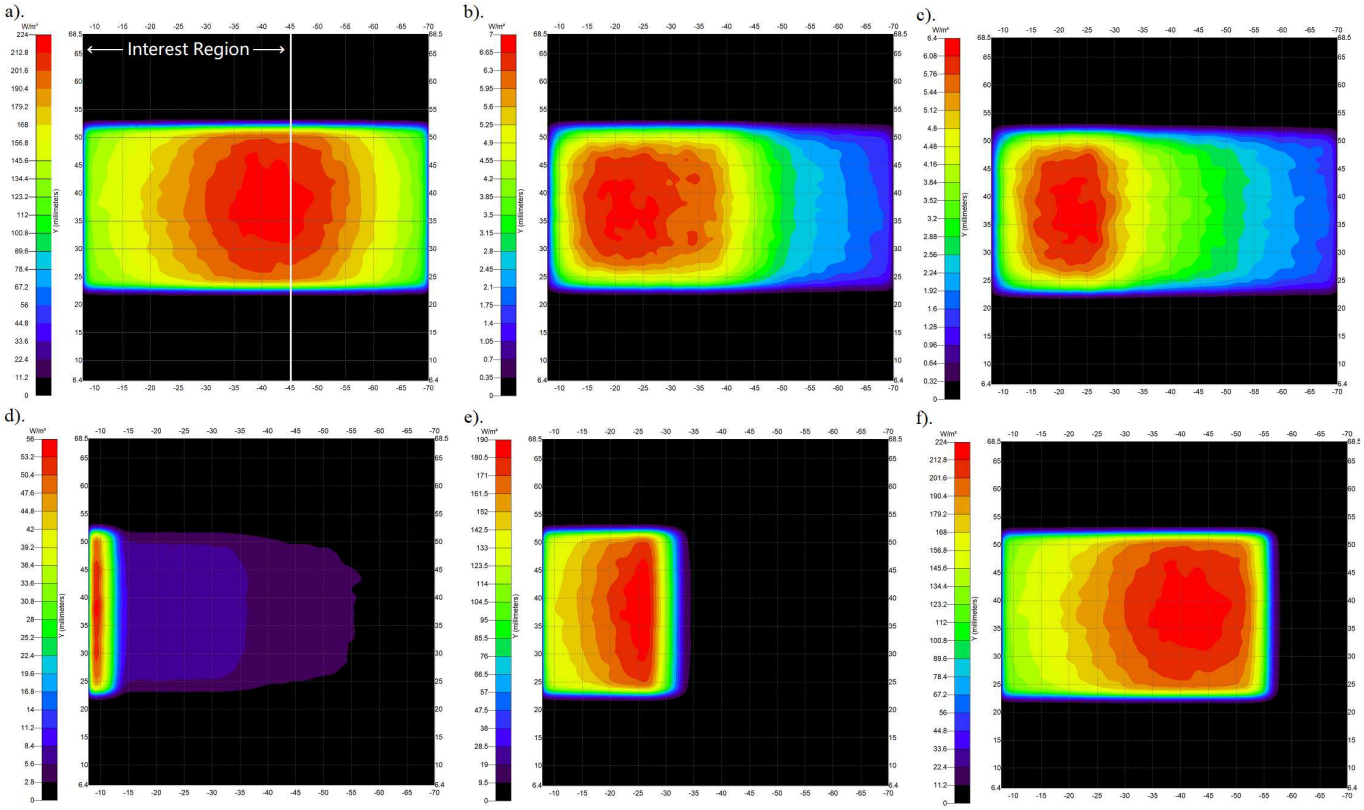

Fig. 6: Luminance maps on the test area for different plate diffuser angles relative to the light source for a) no diffuser, b) 30-degree diffuser, c) 45-degree diffuser, d) 60-degree diffuser, e) 90-degree diffuser, and f) 135-degree diffuser.

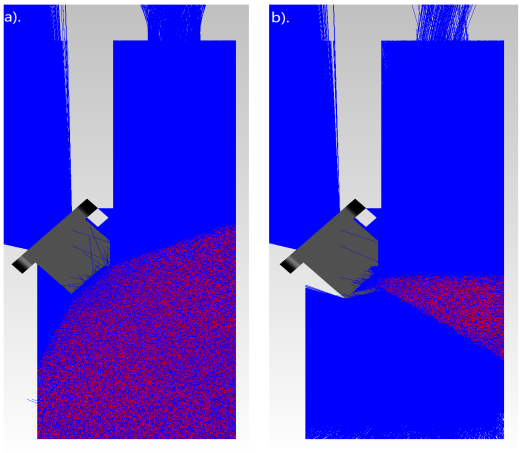

Fig. 7: Optical path comparison a) without diffuser, b) with 30-degree diffuser.

represent the acid, neutral, and alkaline conditions for the solutions, with pH values of 4.01, 7.0, and 9.22, respectively.

Table 2 presents the results of a sample test conducted on three different pH buffer solutions. The findings indicate that, despite the regression model encompassing a smaller pH interval than the classification model, the latter exhibits greater stability in real-world scenarios. This is because the classifi-

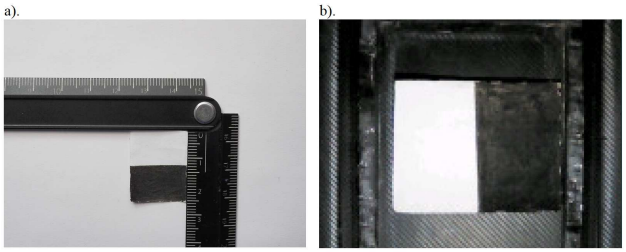

Fig. 8: Calibration reference tape a) geometrical characteristics of the reference tape, b) reference tape utilised condition.

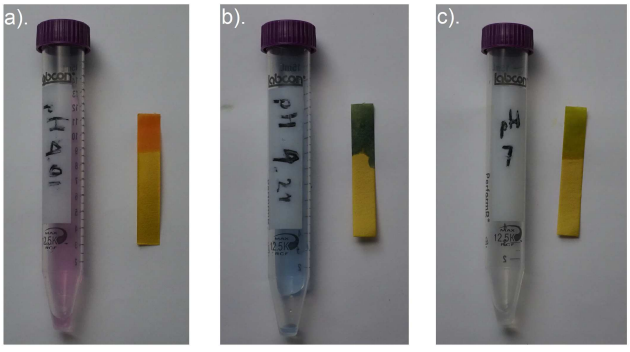

Fig. 9: Three pH buffer solutions for different pH tests a) pH=4.01, b) pH=7.0, c) pH=9.21.

cation model is trained on a group of colour data for each pH value, thereby enabling it to represent a more stable pH result in the presence of greater input colour tolerance. However, the regression model lacks the error tolerance for the input colour data, coupled with the trained model exhibiting a tendency towards overfitting. This ultimately leads to considerable discrepancies in the pH result. Consequently, the classification model is employed in this study and implemented on the test machine.

TABLE II: pH test result for three buffer solutions

Buffer solution pH	**Detect pH**		**Calibrated pH**	
	SVM	***LR***	***SVM***	***LR***
4.01	5	1.51	5	-12.73
7.00	7	5.21	7	-6.63
9.21	8	49.07	9	44.14

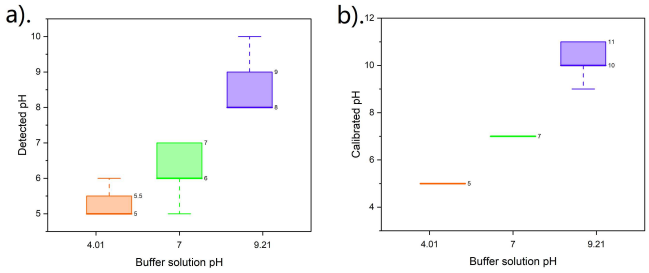

Fig. 10: Testing result analysis a) detected pH value without the colour calibration process, b) detected pH value with the colour calibration process.

In a real test, 20 samples of pH indicator paper are used for each pH buffer solution, with the pH value being determined by the test machine. Figure 10 illustrates the analysis of the test results. A comparison of the pH results obtained without calibration (Figure 10a) and with calibration (Figure 10b) demonstrates that the latter is more stable, which is due to the introduction of the calibration process into the system. Consequently, the calibration algorithm has been demonstrated to reduce the error in the machine vision process at the colour-detecting stage. Nevertheless, Figure 10 also indicates that the pH detection value differs from the actual pH value. This discrepancy can be attributed to the dataset employed for training the machine learning model and the different type of pH indicator used in testing. Despite the implementation of several corrective measures to address the disparity in data caused by the disparate test conditions between the test machine employed in this study and the test machine utilized in the dataset, the discrepancy in the detection of colour persists. Consequently, the inevitable error pH result is obtained by the machine learning model, which has been trained on a different and insufficient dataset.

VI. Conclusion

The present study examines a method that employs machine learning and pH indicator paper to analyse the solution pH value. A test machine has been constructed to provide a controllable environment for machine vision processes that detect the colour of the pH indicator paper. In this study, several machine learning models were trained using different machine learning algorithms and datasets, of which the best performing for the SVM model achieved an accuracy of 0.95, and for the RT model achieved an RMSE and R^2 of 0.32 and 0.99 respectively. The experimental results demonstrate that, despite the regression model exhibiting smaller pH intervals than the classification model, the latter is more stable in real-world testing. This demonstrates the potential for regression model to achieve more stable pH testing. Future study should consider the creation of an original pH-colour dataset with the aim of improving the machine learning model.

References

[1] D. Webster, "PH–PRINCIPLES AND MEASUREMENT," in Encyclopedia of Food Sciences and Nutrition, Second., B. Caballero, Ed., 2003, pp. 4501–4507. doi: https://doi.org/10.1016/B0-12-227055-X/00913-5.

[2] L. E. V. Salgado and C. Vargas-Hernández, "Spectrophotometric Determination of the pKa, Isosbestic Point and Equation of Absorbance vs. pH for a Universal pH Indicator," Am J Analyt Chem, vol. 05, no. 17, pp. 1290–1301, 2014, doi: 10.4236/ajac.2014.517135.

[3] J. T. Woods and M. G. Mellon, "A Spectrophotometric Study of Universal Indicators," J Phys Chem, vol. 45, no. 2, pp. 313–321, 1941, doi: 10.1021/j150407a015.

[4] M. M. Elsenety, M. B. I. Mohamed, M. E. Sultan, and B. A. Elsayed, "Facile and highly precise pH-value estimation using common pH paper based on machine learning techniques and supported mobile devices," Sci Rep, vol. 12, no. 22584, Dec. 2022, doi: 10.1038/s41598-022-27054-5.

[5] S. K. Sunori, S. Kumar, B. Anandapriya, S. L. Nesamani, S. Maurya, and M. K. Singh, "Machine Learning Based Prediction of Soil pH," in Proceedings of the 5th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2021, Coimbatore, India: IEEE, 2021, pp. 884–889. doi: 10.1109/ICECA52323.2021.9675926.

[6] A. I. Gomez, J. M. Gutiérrez, and R. Muñoz, "Design of an Optical Portable System for Detection of pH," in Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), Porto, Portugal: IEEE, 2018, pp. 1–6. doi: 10.1109/GMEPE-PAHCE.2018.8400745.

[7] Ali. Y. Mutlu, V. Kiliç, Gizem. K. Özdemir, A. Bayram, N. Horzum, and Mehmet. E. Solmaz, "Smartphone-based colorimetric detection: Via machine learning," Analyst, vol. 142, no. 13, pp. 2434–2441, Jul. 2017, doi: 10.1039/c7an00741h.

[8] A. Dhar, R. Mehta, and M. Sit, "Colorimetric Detection of pH Strips," Dec. 2017.

[9] M. Hoque Tania, K. T. Lwin, A. M. Shabut, M. Najlah, J. Chin, and M. A. Hossain, "Intelligent image-based colourimetric tests using machine learning framework for lateral flow assays," Expert Syst Appl, vol. 139, Jan. 2020, doi: 10.1016/j.eswa.2019.112843.

[10] R. B. G. Luta, R. G. Baldovino, and N. T. Bugtai, "Multi-label Classification of pH Levels using Support Vector Machines," in 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines: IEEE, 2018, pp. 1–4. doi: 10.1109/HNICEM.2018.8666299.

[11] R. Rojan, "pH-recognition." Accessed: May 18, 2024. [Online]. Available: https://www.kaggle.com/datasets/robjan/ph-recognition

[12] Otsu N, "A Threshold Selection Method from Gray-Level Histograms," IEEE Trans Syst Man Cybern, vol. 9, no. 1, pp. 62–66, 1979.

[13] Zhao Yunfeng, "Image data collection, processing, storage, and their application in smartphone food analysis," Queen's university of Belfast, 2022.

[14] P.-L. Chen et al., "The Illumination Uniformity Study of Diffuser Plates for Reflective LED Desk Lighting," in 9th IEEE International Conference on Nano/Micro Engineered and Molecular Systems, Waikiki Beach, HI, USA: IEEE, 2014, pp. 525–529. doi: 10.1109/NEMS.2014.6908864.

[15] J. Il Hong and B. Y. Chang, "Development of the smartphone-based colorimetry for multi-analyte sensing arrays," Lab Chip, vol. 14, no. 10, pp. 1725–1732, 2014, doi: 10.1039/c3lc51451j.