Determination of Soluble Copper in Water through Tannic Reaction Analysis using an Optical Color Sensor and Machine Learning

Jan Fern Historillo[1,2], Gerhard P. Tan[2,3], Robert G. De Luna[2,3]
[1]College of Industrial Technology, Marinduque State College, Marinduque, Philippines
[2]Graduate School, Polytechnic University of the Philippines, Sta. Mesa, Manila
[3]Gokongwei College of Engineering, De La Salle University, Manila
*janfernhistorillo@iskolarngbayan.pup.edu.ph, gptan@pup.edu.ph, rgdeluna@pup.edu.ph

Abstract: Dissolved copper in water is considered before it can be used for any purpose, especially in the case of Marinduque, Philippines where mine tailing containing high concentrations of the metal was spilled to various water resources. This study aimed to use an optical fiber amplifier to record the color reaction being produced by soluble copper and develop a machine-learning model that determines the level of soluble metals in water. The research utilized a prototype and AI development framework. Preparation of the image processing device and fiber optic sensor (BV501s), the preliminary gathering of baseline data, system modeling & development, and System Evaluation was conducted. The BV501s optical fiber sensor was used as it provides RGB equivalents of tannic reaction to soluble copper. There was a total of 33 samples each with its segmented RGB equivalents. The model for identification of the amount of soluble metal was developed using two (2) models such as multiple regression analysis and support vector regression. Using the applied methodology, copper amounts can be determined in water using regression models. The amount of soluble copper can be identified more accurately using the support vector regression since it yields a higher r2 score compared to the multiple linear regression model using k-fold cross-validation. Integration of the developed model to hardware is also suggested as and comparison of the developed model. Develop a classification model for the potability of the water samples based on the soluble water samples.

Keywords: *Copper, Optical Fiber Sensing, Machine Learning, Regression Models, Water Quality*

I. INTRODUCTION

One of the three (3) important factors before water can be used is the dissolved matter in it. For water to be used for drinking or domestic premises, bacteriological, physical, and dissolved matters are highly considered. Acceptable levels of various parameters in the said consideration were issued by the World Health Organization. Exceeding such limits and acceptable standards leads to various health conditions and impacts on living matters in the environment. [1]

In the province of Marinduque, high concentrations of soluble metals are in water and soil, particularly Copper and Iron. In 1996 and 1998, two major tragic mining incidents in the country happened that contaminated various waterways and sources in the province. This resulted in serious health problems for the people within the area and resulted in decreased crop production even within 20 years after the disaster. Water resources have been heavily contaminated and even aquatic resources where the rivers lead declined. [2] [3]

As an island province access to laboratories to assess soluble metals in water has been a challenge. To address these challenges, the sap of Physic *Jatropha curcas* is mixed with sample water to test if it's potable or not. The color reaction produced by the tannic acid and soluble metal is used to determine the water's potability. [4] [5] [6] This is illustrated in Figure 1.

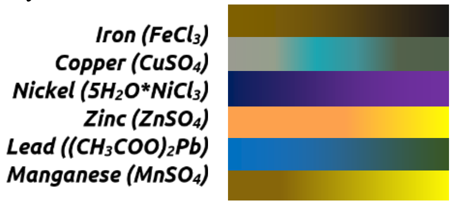


***Figure 1**. Color Reaction of Tannins to Soluble Copper [5]*

As shown in the preceding figure, soluble copper when reacting with tannins produces a light blue shade to a dark green gradient. Various studies established that this color reaction is caused by the tannic effect on the pigment structure of soluble metals. In the study of Historillo in 2022, the determination of soluble copper in water was established using a camera and color and imaging approach. The study analyzed Red-Green-Blue (RGB) segments of the color reaction. A high-resolution camera was installed in an image processing box to record the color reaction. The recorded color reaction is compared to a color chart. The developed system was found to be 95% accurate compared to the result produced in the laboratory. [7]

Color segmentation is a technique that separates the red, green, and blue components of a given color. This technique enables the user to have numerical data that can be used for machine learning and the development of algorithms. [8][9] It has a wide array of applications such as counting, inspection, and most of all classification of materials. The basic principle relies on positioning the color detection device at a certain distance. Using the optical sensor, the color is absorbed and compared to certain point values such as their Red-Green-Blue counterparts. [10]

Research reveals that optical fiber sensor has higher efficiency and accuracy in light-related applications due to their characteristics such as characteristics of reflectivity and refractive properties, operation at the wide spectrum of the light source, and low power consumption. [8] Its ability to be adapted to various applications such as manufacturing, and color detection makes it more viable than other sensing devices. [9] It can detect various light spectrums by division of the wavelength. This function is the same with its capability to detect light by emitting light and measuring the value of the reflected light waves and segmenting its Red-Green-Blue equivalents. [10] [11][12] Useful techniques and approaches combine various algorithms and several devices. Such algorithms include (but are not limited to) regression and classification techniques. Combining segmentation of different color spaces yields more accurate segmentation results than segmentation of a single-color space. K-means clustering and Effective robust kernelized fuzzy c-means are used to segment the images. [13][14][15]

Given these concepts, this study aimed to use a color-sensing optical fiber amplifier to record the tannic reaction of soluble copper and develop a machine-learning model that determines the level of soluble copper in water. Specifically, this research aims to ascertain the accuracy of the developed system through cross and hold-out validations.

This study is expected to benefit communities where access to laboratories is a challenge. The study offers a new technique and approach to determining soluble copper in water. The study also shows another application of optical sensors. However, this study is only limited to determining dissolved metal ($CuSo^{4+}$) in water from 0.25 to 5 ppm (parts per million / grams/Liter). The study was conducted from January to February 2023. Python language programming shall be used particularly the sci-kit learn and related libraries to create the system.

II. Methodology

This research utilized a prototype and AI development framework. The prototype and system design were based on existing data which is a product of conducted experiments. The said design was deemed fit for this research as it will provide the necessary data for the objectives of this research.

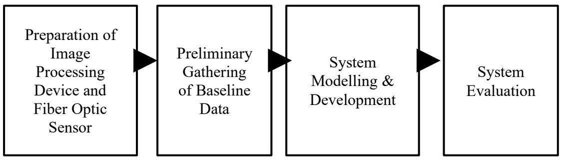
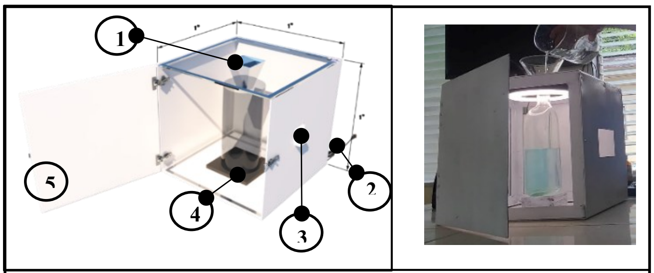


Figure 2. *Research Theoretical Framework*

As shown in Figure 2, there were four (4) necessary phases to achieve the objectives of this study such as planning and acquisition of materials needed, preliminary data gathering of baseline data, system modeling and development, and evaluation.

A. Preparation of Image Processing Device and Fiber Optic Sensor

Research planning and acquisition of materials necessary to develop the system are to be accomplished first. Necessary readings and consultation with subject matter experts shall be done as well. A review of existing techniques for determining copper in water was also included in this stage. An image processing box with a sufficient proper lighting system was prepared including the optical fiber sensor as shown in Figures 3 and 4. The fabricated image acquisition material was prepared to set the area of interest during the image acquisition process and maintain constant lighting throughout the acquisition process. The optical fiber sensor used was commercially available with model number BV501s which is widely used in manufacturing and production processes. It has an output display that shows the numerical values of the RGB components of the color being read and interpreted. The two pieces of hardware were integrated.



1: Tannic Acid Slot.2: LED and Optical Fiber Cable: 3: Optical Fiber Slot 4: Adjustable Platform 5: Box Door

Figure 3. *Image Capturing Box for Color Reaction Analysis*

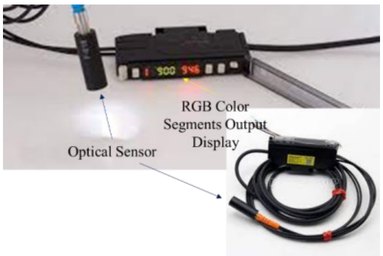


Figure 4. *BV501s Optical Fiber Sensor*

B. Preliminary Gathering of Baseline Data

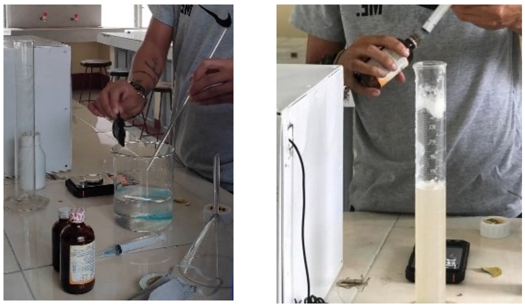


Figure 5. *Preparation of Tannic Acid Solution and Water Samples*

Figure 5 shows the preparation of the water sample solutions and the tannic acid from *Jatropha curcas.* The replicates were made by dissolving specific amounts of soluble copper varying from 0.25 ppm to 5 ppm to 100 mL of pure and distilled water. 30 $^{v}/_{v}$ tannic acid was dissolved in the prepared metal solutions. This volume-per-volume concentration was based on the study of Historillo (2015) where significant color change can be observed. [4] The color reaction was listed together with the time for a noticeable change to happen. Since the optical fiber sensor provides RGB components of the color reaction, the actual color reaction was acquired using the MatLab color and imaging toolbox. The color reactions were determined based on their Red-Green-Blue Components. The data set that was used in this study was based on the three (3) replicates of color reaction for soluble copper.

Features for Determination of Amount of Soluble Copper

The following are the features that were used to develop the model. These features were established as observed from the actual color reaction of tannic acid to dissolved copper.

A. Red Component of Tannic Color Reaction 1
B. Green Component of Tannic Color Reaction 1
C. Blue Component of Tannic Color Reaction 1
D. Red Component of Tannic Color Reaction 2
E. Green Component of Tannic Color Reaction 2
F. Blue Component of Tannic Color Reaction 2
G. Time Elapsed for observed full-color Reaction.
H. Amount of Tannic Acid Dispensed
I. Amount of Dissolved Soluble Copper

There was a total of 33 samples each with its segmented RGB equivalents. Since the color reaction patterns mostly consist of a change in color reaction, two (2) specific points were selected. These points were selected uniformly for all the samples using the Color Segmentation feature. As established by earlier studies, the time it took for color reaction to happen is directly correlated with the amount of dissolved metal. A soluble amount of copper can be determined relative to time and the color gradient produced.

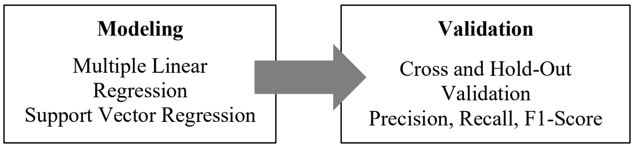


Figure 6. *System Development, Modeling and Validation*

Once the data set was established using the reverse color segmentation was conducted using MatLab, it was saved as a Comma-Delimited File (CSV). The data set was checked for missing data and non-uniform entries using the Pandas library of Python.

C. Systems Development

For the models to be developed, Figure 6 shows the methods and techniques used to create, validate and optimize the model. The model for identification of the amount of soluble metal was developed using two (2) models such as multiple regression analysis and support vector regression.

D. System Evaluation

For the regression models, k-10-fold cross-validation was used to determine the accuracy of the developed system comprised of mean absolute error, mean squared error, and r^2 error. Hold-out validation was also conducted to determine whether the developed system was fit.

III. Results and Discussion

The following results are hereby presented based on the used methodology in developing the classification and regression models.

A. Generated Red-Green-Blue Components of the Fiber Optic Sensor

Table 1. *Summary of the Result of the Experimental Setup and the RGB Equivalents as generated by the Optical Fiber Sensor*

Replicate No.	Amount of Tannic Acid	Time elapsed	Dissolved Copper (in ppm)	Actual First Color Generated	Red Equivalent 1	Green Equivalent 1	Blue Equivalent 1	Actual Second Color Generated	Red Equivalent 2	Green Equivalent 2	Blue Equivalent 2
1	30 mL	37	0.25		255	255	255		243	255	230
2	30 mL	38	0.25		255	255	255		244	256	231
3	30 mL	35	0.25		253	253	253		241	253	228
1	30 mL	29	0.5		255	235	37		205	228	190
2	30 mL	28	0.5		254	234	36		204	227	189
3	30 mL	27	0.5		253	233	35		203	226	188
1	30 mL	26	1.0		221	255	246		161	181	150
2	30 mL	25	1.0		220	254	245		160	180	149
3	30 mL	27	1.0		222	256	247		162	182	151
1	30 mL	26	1.5		200	248	232		148	168	138
2	30 mL	25	1.5		199	247	231		147	167	137
3	30 mL	27	1.5		201	249	233		149	169	139
1	30 mL	24	2.0		72	197	190		63	85	52
2	30 mL	22	2.0		73	198	191		64	86	53
3	30 mL	24	2.0		73	198	191		64	86	53
1	30 mL	22	2.5		122	192	174		59	83	51
2	30 mL	20	2.5		121	191	173		58	82	50
3	30 mL	22	2.5		123	193	175		60	84	52
1	30 mL	21	3.0		136	143	119		40	53	34
2	30 mL	19	3.0		135	142	118		39	52	33
3	30 mL	21	3.0		134	141	117		38	51	32
1	30 mL	20	3.5		96	112	85		30	40	26
2	30 mL	18	3.5		95	111	84		29	39	25
3	30 mL	20	3.5		94	110	83		28	38	24
1	30 mL	19	4.0		93	105	80		29	38	26
2	30 mL	17	4.0		93	105	80		29	38	26
3	30 mL	19	4.0		93	105	80		29	38	26
1	30 mL	18	4.5		88	90	74		26	34	24
2	30 mL	16	4.5		88	90	74		26	34	24
3	30 mL	18	4.5		88	90	74		26	34	24
1	30 mL	17	5.0		42	53	40		30	40	26
2	30 mL	15	5.0		42	53	40		30	40	26
3	30 mL	17	5.0		42	53	40		30	40	26

As the data set used in this research is a controlled experimental setup, data wrangling was done using the Pandas library of Python.

All of the data used except for the soluble copper present was classified as an integer.

Table 1 shows the summary of the result of experimental setups and the generated values for each replicate. For each dissolved metal, there were 33 samples used with varying amounts of metals. There were no outliers in the data set used. The average time for the observable tannic reaction to happen is 23 seconds.

The RGB color component ranges from 0 to 255 depending on the color of the tannic acid reaction. The color gradient changes from yellow to dark green depending on the amount of soluble metal present. This is in congruence with the findings of Marquez, (2008), Historillo (2015), and Historillo (2022) that for soluble copper, the chromaticity index of the color reaction changes relative to the amount of metal present in the sample. [4] [7] As established the higher the concentration of soluble metal, the higher the chromaticity index or the darker the color being produced.

Using Pearson correlation, data revealed that the relationship between the time elapsed and the amount of soluble metal is -0.9068. This supports the established studies that the higher the amount of tannic acid, the lesser the time for an observable reaction to occur. The regression chart is presented in Figure 6.

B. *Identification of Amount of Soluble Metals in Water*

Using two regression techniques such as multiple linear and support vector regression, identification of the amounts of soluble metals present in water were developed. Table 2 presents the cross and hold-out validation scores for each model. It is noticeable that for both models the validation measures are the same with the r^2 value of 0.86.

Table 2. *Validation of the Developed Regression Models*

	Cross Validation			Hold Out Validation		
Multiple Linear Regression						
	Mean Absolute Error	Mean Squared Error	R^2	Mean Absolute Error	Mean Squared Error	R^2
Average	0.244	0.083	0.914	0.251	0.092	0.988
Deviation	0.078	0.047	0.03	Explained Variance Score (EVS)		0.988
Support Vector Regression						
	Mean Absolute Error	Mean Squared Error	R^2	Mean Absolute Error	Mean Squared Error	R^2
Average	0.238	0.077	0.748	0.121	0.015	0.991
Deviation	0.079	0.044	0.606	Explained Variance Score (EVS)		0.993

It can be observed that the model developed using support vector regression outperforms the other model that used multiple linear regression in terms of the mean absolute errors and mean squared errors. The r^2 value (for hold-out validation) for the support vector regression was computed at 0.991 while the other model was only at 0.988. The mean absolute error and mean squared error for support vector regression are also lesser than multiple linear regression values, thus having a higher accuracy when using the K-Fold validation.

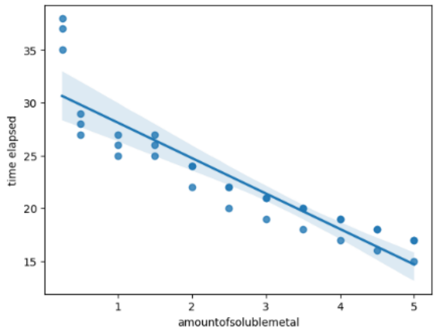


Figure 6. *Regression Chart of Amount of Soluble Metal and Time Elapsed*

Relative to explained variance score, support vector regression has a score closer to 1 compared to the multiple linear regression model. This supports the support vector regression in terms of performance in detecting soluble copper in water. In determining amounts of soluble copper in water, support vector regression provides a more accurate output based on the RGB color equivalents of the color reaction and the time it takes to produce a reaction.

Upon using BV501 optical fiber sensor, RGB color segmentation approach was found to be useful in determining soluble copper in water. Compared to the previous studies established [7] [12], the use of the optical color sensor provided a direct Red-Green-Blue equivalents rather than segmenting a given color. This technique provides a better and easier approach as numerical equivalents are directly given and can be used in the developed regression models.

IV. Conclusion

Using the color-sensing optical fiber amplifier to record the tannic reaction of soluble copper and regression models, a machine learning model was developed that determines the level of soluble copper in water. Results reveal that the amount of soluble copper can be identified more accurately using the support vector regression since it yields a higher r^2 score compared to the multiple linear regression model using k-fold cross-validation.

Integration of the developed model to hardware is also suggested as and comparison of the developed model. Develop a classification model for the potability of the water samples based on the soluble water samples.

V. REFERENCES

[1] World Health Organization, Guidelines for Drinking Water Quality, Interligar, Brazil: World Health Organization, 2022.

[2] E. Buella-Mandia, "Ex-Post Environmental Assessment of Mined-Out Areas In Boac, Marinduque, Philippines: Two Decades After The Disaster," School of Environmental Science and Management, University of the Philippines. Los Baños Laguna, 2023.

[3] J. Cruz, J. Baquiran, J. Suan, I. Villalba and J. Cruz, "Project Daluyan: Water Quality, Hydrology, and Hydraulic Analyses of Boac River, Marinduque," Marinduque Provincial Government Environment and Natural Resources Office, College of Engineering, University of the Philippines Diliman, Quezon City, 2020.

[4] J. F. Historillo, Microprocessor-Based Tannic Acid Dispenser: A Tool for Color Reaction Analysis of Deep-Well Waters in Marinduque, Boac, Marinduque: Marinduque State College, 2015.

[5] M. A. Marquez, A. R. Malapad and R. A. Chi, "The effectiveness of (*Jatropa curcas*) sap as metallic ion indicator in water through color reaction," in Intel Philippine Science and Engineering Fair, Tagaytay, 2008.

[6] R. J. Marquez, J. F. Historillo and J. R. Padolina, "*Jatropa curcas* Sap and Paper Strips as Heavy Metal Indicator in Water," in Intel Philippine Science and Engineering Fair, Clark City, Subic, Pampanga, 2009.

[7] J. F. Historillo, "Color and Image Analysis Approach in Determination of Soluble Copper in Water Using Tannic Reaction Analysis," in 5th International Conference in Intelligent Computing and Optimization, Thailand, 2022.

[8] N. Guo, J. Liu, Q. He, R. Zhou and H. Yuan, "Design of 3D-Printed Electronic Fiber Optic Sensor to Detect Rhodamine B Reagent: An Initiation to Potential Virus Detection," Biomimetics, vol. 7, no. 94, 2022.

[9] R. Gothwal, S. Gupta, D. Gupta and A. K. Dahiya, "Color image segmentation algorithm based on RGB channels," *Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization*, Noida, India, 2014, pp. 1-5, doi: 10.1109/ICRITO.2014.7014669.

[10] National Instruments Corp. Color Segmentation. https://www.ni.com/docs/en-US/bundle/ni-vision-concepts-help/page/color_segmentation.html#:~:text=Color%20segmentation%20compares%20the%20color,of%20interest%20from%20background%20clutter. March 2023

[11] Keyence, "Introductory Guide to Sensors," Keyence Corporation, 2023. [Online]. Available: https://www.keyence.com/ss/products/sensor/sensorbasics/color/feature/.

[12] P. Cao., Y. Zhu., W. Zhao, S. Liu "Chromaticity Measurement Based on the Image Method and Its Application in Water Quality Detection," MDPI's Monitoring, Modelling and Management of Water Quality, vol. 11, no. 11, 2019.

[13] C. Mythili. V. Kavitha. Color Image Segmentation using ERKFCM. *International Journal of Computer Applications.* Vol. 41. Issue 20. March 2012

[14] I. A. Grout, M. Zaidi, K. L. Sterckx and A. K. A'ain, "RGB LED Driver Circuit Design for an Optical Fiber Sensor System," ECTI Transactions on Computer and Information Technology (ECTI-CIT), vol. 11, no. 2, pp. 163-177, 2017.

[15] C. Pendão and I. Silva, "Optical Fiber Sensors and Sensing Networks: Overview of the Main Principles and Applications," October 2022. [Online]. Available: https://doi.org/10.3390/s22197554. [Accessed January 2023].

[16] R. Gothwal, S. Gupta, D. Gupta and A. K. Dahiya, "Color image segmentation algorithm based on RGB channels," *Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization*, Noida, India, 2014, pp. 1-5, doi: 10.1109/ICRITO.2014.7014669.