Supporting Information

An Ensemble Model of Machine Learning Regression Techniques and Color Spaces Integrated to a Color Sensor: Application to Color-Changing Biochemical Assays

Min Joh[a], Surjith Kumaran[a], Younseo Shin[a], Hyunji Cha[a], Euna Oh[a], Kyu Hyoung Lee[b], and Hyo-Jick Choi[a,*]

[a]Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
[b]Department of Materials Science and Engineering, Yonsei University, Seoul 03722, Republic of Korea

[*] Corresponding author: hyojick@ualberta.ca (H.-J.C).

Abstract: Non-destructive color sensors are widely applied for rapid analysis of various biological and healthcare point-of-care applications. However, existing RGB-based color sensor systems, relying on the conversion to human-perceptible color spaces like HSL or HSV, exhibit limitations compared to spectroscopic methods. The integration of machine learning (ML) techniques presents an opportunity to enhance data analysis and interpretation, enabling insights discovery, prediction, process automation, and decision-making. In this study, we utilized four different regression models integrated with an RGB sensor for colorimetric analysis. Colorimetric protein concentration assays, such as BCA and Bradford analysis, were chosen as model studies to evaluate the performance of the ML-based color sensor. Leveraging regression models, the sensor effectively interprets and processes color data, facilitating precision color detection and analysis. Furthermore, the incorporation of diverse color spaces enhances the sensor's adaptability to various color perception models, promising precise measurement, and analysis capabilities for a range of applications.

Supporting Information

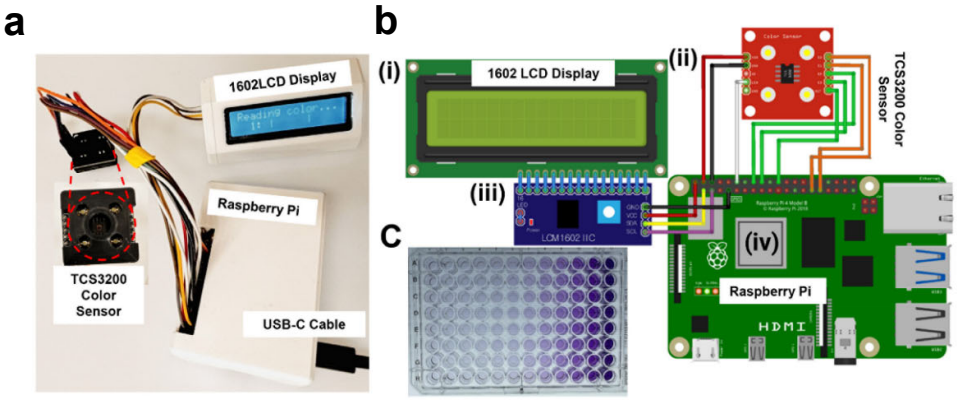

Figure S1. Color sensor for the determination of protein concentration: Photo of the fabricated prototype (a), circuit diagram of the LCD display (b-i), RGB sensor (b-ii), LCD circuit module (b-iii), Raspberry Pi minicomputer (b-iv), and 96-well plate BCA Assay used in the commercial plate reader (c).

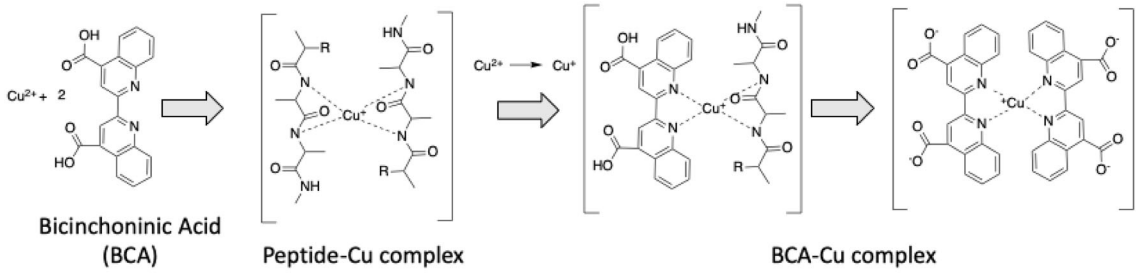

Figure S2. Mechanism of BCA assay for protein quantification.

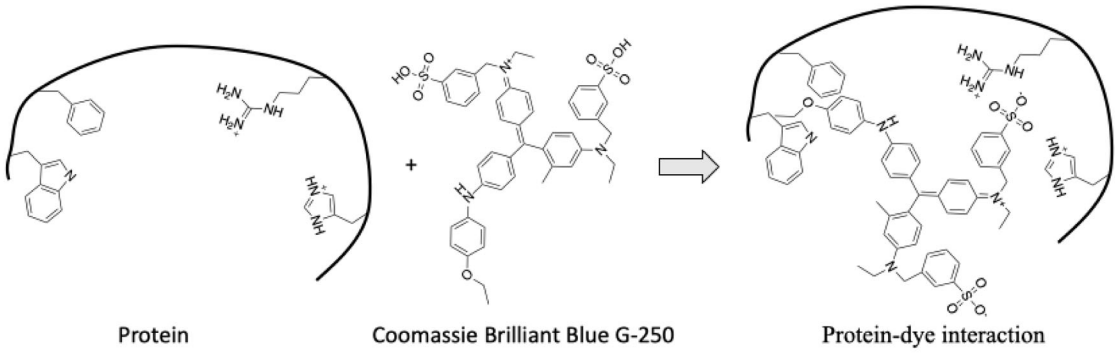

Figure S3. Mechanism of Bradford assay for protein quantification.

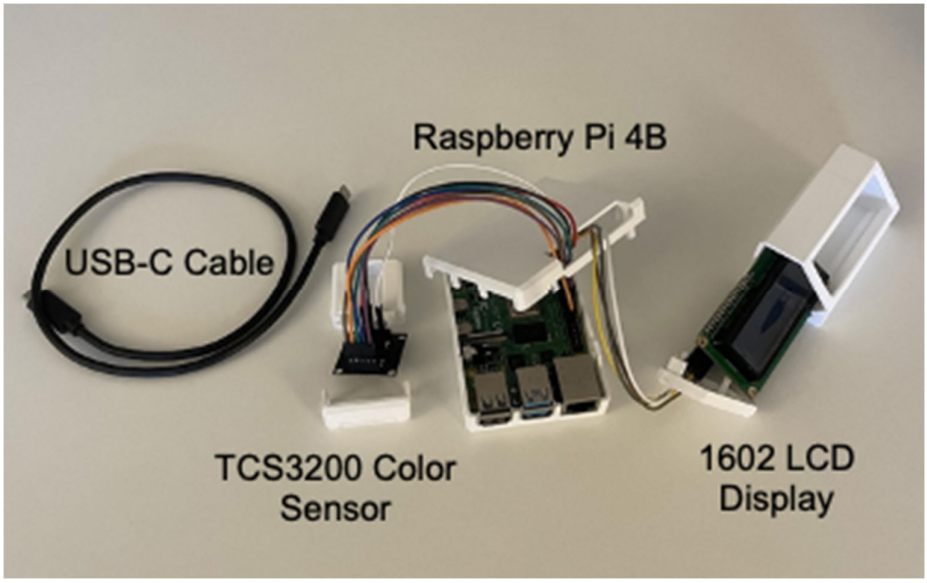

Figure S4. Image of manufactured prototype of the color sensor with labels

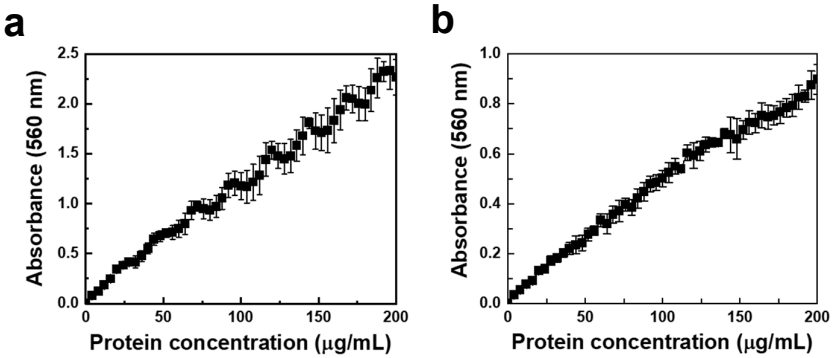

Figure S5. Protein quantitation curve of absorbance at 560 nm vs protein concentration (0-200 μg/mL) in H_2O of BCA assay (a), Bradford assay (b).

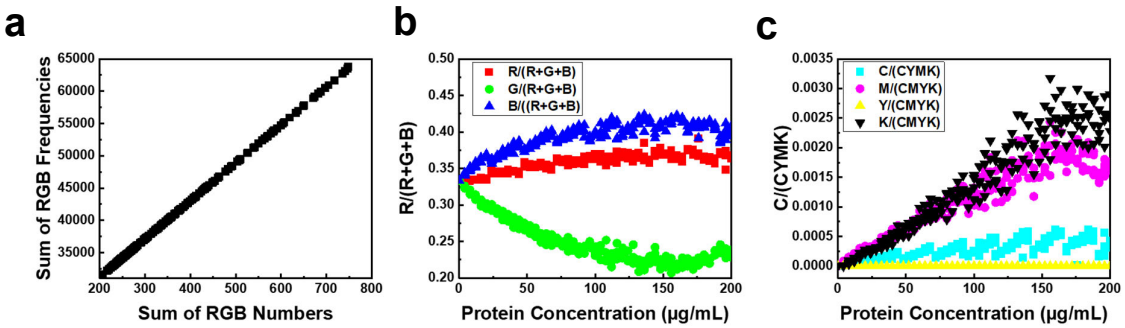

Figure S6. RGB color component values vs. frequency of color sensor output values (a), ratio of individual colors (R, G, B) to the sum of RGB with increasing protein concentration (b), and ratio of individual C, M, Y, K colors to the sum of CMYK (c) with increasing protein concentration for the BCA assay test.

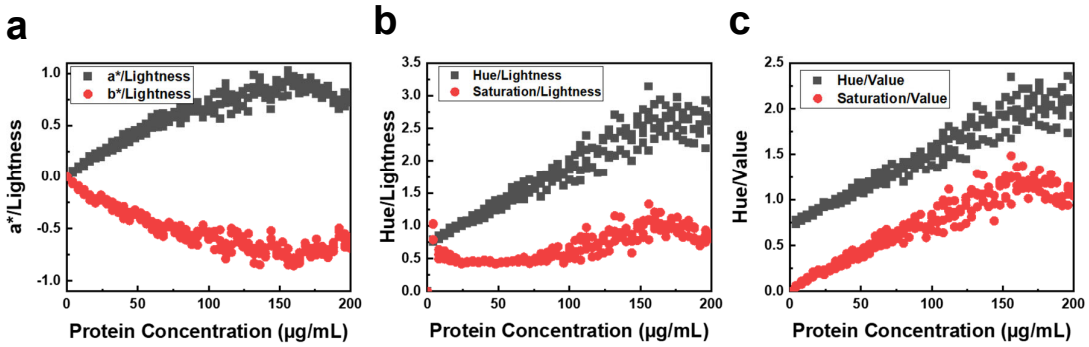

Figure S7. Calibration curves for CIELAB, HSL and HSV values: a* value (green) and b* value (blue) over lightness in CELAB (a) hue and saturation over lightness vs. increasing protein concentration (b), and corresponding hue and saturation over value in HSV (c), for 0-200 µg/mL of BSA protein samples in the Bradford assay

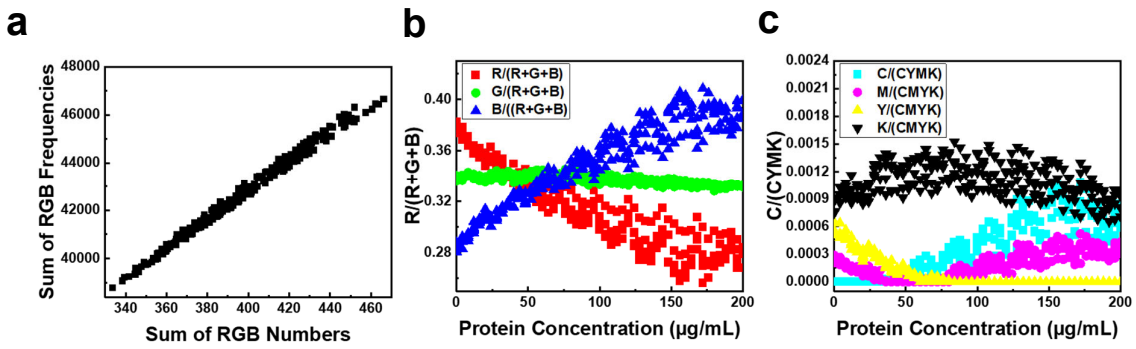

Figure S8. Sum of RGB color component values vs. frequency of color sensor outputs (a), ratio of individual R, G, B colors to the sum of RGB (b), and ratio of individual C, M, Y, K colors to the sum of CMYK (c) with increasing protein concentration for the Bradford assay test

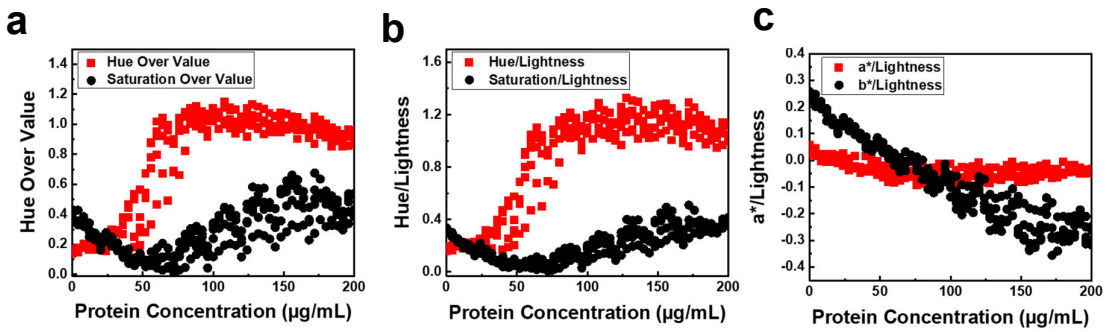

Figure S9. Calibration curves for HSL values: hue and saturation over lightness vs. increase in protein concentration (a), corresponding hue and saturation over value in HSV (b), and the a* value (green) and b* value (blue) over lightness in CIELAB (c) for 0-200 µg/mL BSA protein samples for the Bradford assay.

Supporting Information

Table S1. Summary of Regression model error matrices of various color models

HSL

Regression model	MAE	MSE	RMSE	R^2
Random forest	7.252699	93.10432	9.649048	0.959404
Gradient boosting	7.619387	96.647119	9.830927	0.957859
SVM	8.611231	111.836115	10.575260	0.951237
MLP	29.482344	1260.028438	35.496879	0.450596

RGB

Regression model	MAE	MSE	RMSE	R^2
Random forest	8.100157	106.597984	10.324630	0.953521
Gradient boosting	8.945299	128.054205	11.316104	0.944165
SVM	7.309625	80.945129	8.996951	0.964706
MLP	11.024903	180.368798	13.430145	0.921355

CIELAB

Regression model	MAE	MSE	RMSE	R^2
Random forest	7.608305	94.469397	9.719537	0.958809
Gradient boosting	7.603124	92.615470	9.623693	0.959617
SVM	8.690862	147.183263	12.131911	0.935824
MLP	7.366562	87.956671	9.378522	0.961649

HSV

Regression model	MAE	MSE	RMSE	R^2
Random forest	7.257158	85.341868	9.238066	0.962789
Gradient boosting	7.977991	102.496236	10.124042	0.955309
SVM	8.200111	106.195190	10.305105	0.953696
MLP	25.362530	980.959484	31.320273	0.572277

CMYK

Regression model	MAE	MSE	RMSE	R^2
Random forest	8.192314	115.362849	10.740710	0.949699
Gradient boosting	7.449949	91.161053	9.547830	0.960251
SVM	7.756273	90.476935	9.511936	0.960550
MLP	29.315475	1223.267778	34.975245	0.466624

Table S2. Summary of previous color sensor and regression model work reported

Detector Device	Target Object	Colorimetry Model	ML Model	Performance	Reference
Smartphone Camera	pH values from pH strips	Mean RGB	LS-SVM	Accuracy = 100%	1
CMOS camera	Concentration of a laser dye	HSV and RGB	Polynomial regression	Accuracy = 95.5%	2
TCS34725 on RPi4	Concentrations of food dyes	RGB Ratios	Linear Regression	R-square = 99.3%	3
TCS3200 on Arduino	Concentration of cyanide	RGB	Linear Regression	R-square = 98.0%	4

Table S3. Summary of advantages and drawbacks of regression models used in this study

Regression models	Advantages	Drawbacks
Random Forest Regressor (RFR)	An ensemble method used for predicting continuous outcomes by combining the outputs of multiple decision tree regressor, enhancing accuracy and robustness through max voting.[5] This method reduces variance and overfitting, providing stable predictions, particularly advantageous in color sensor data susceptible to environmental noise.[6]	It can be computationally intensive and may not perform well with high-dimensional data.[9]
Gradient Boosting Regressor (GBR)	Combines multiple weak learners (decision trees) into a strong model, improving accuracy. Each tree leans from the errors of the previous one, optimizing predictions iteratively.[9] Additionally, RFR and GBR facilitate easy recalibration and adaptation to different devices and conditions.[7]	Computationally intensive due to the large number of estimators. Prone to overfitting if hyperparameters (like learning rate and number of estimators) are not well-tuned.[8]
Support Vector Regressor (SVR)	Works well for both linear and non-linear relationships due to kernel functions and is highly effective for high-dimensional data.[8]	It is less effective in low-feature classes. SVR struggles in datasets with a small number of features, as its performance relies on having sufficient dimensionality to distinguish patterns.[8]
Multi-Layer Perceptron (MLP)	Excels at identifying complex patterns and non-linear relationships in the data.[8]	Prone to overfitting if not properly regularized and if the data is too simple. Computationally expensive for large datasets due to the complexity of neural network architecture.[8]

Supporting Information

Lysozyme is a widely distributed low-molecular-weight protein with antimicrobial activity. It is commonly found in various species, including viruses, bacteria, fungi, and mammals, and is particularly abundant in bodily fluids such as tears, saliva, milk, and mucus.[10] To further validate the color sensor and ML algorithm, we performed additional protein concentration analysis using colorimetric assays to detect lysozyme and further validate the test.

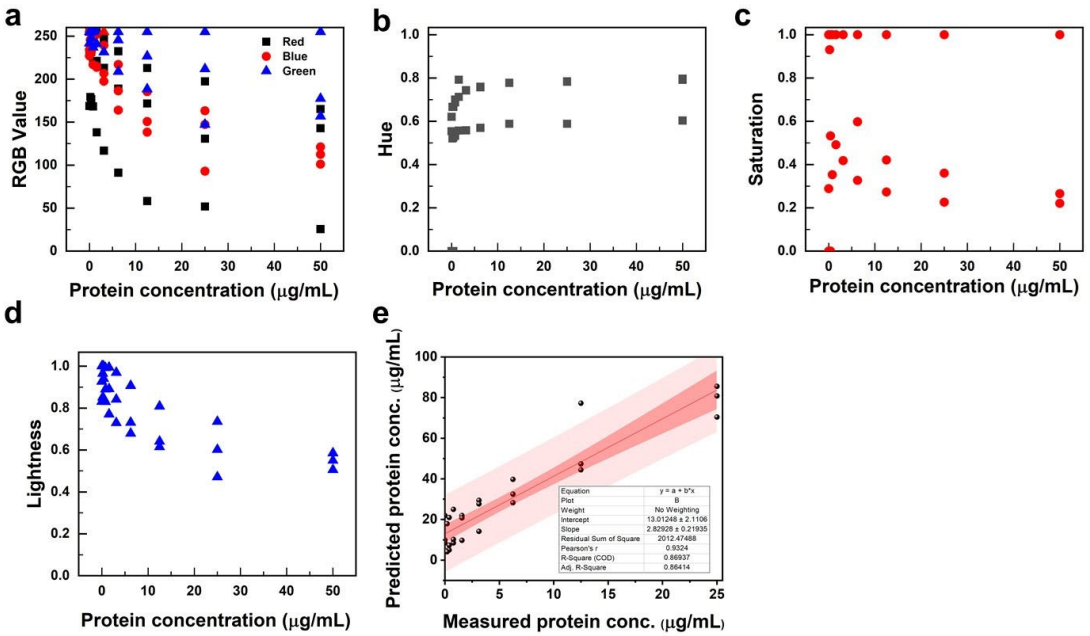

Figure S10. RGB color component values vs. frequency of color sensor output values (a), Hue colors (H) with increasing protein concentration (b), Saturation colors (S) with increasing protein concentration (c), Lightness (L) with increasing protein concentration (d) and a scatterplot illustrating prediction vs measured protein concentration of the lysozyme.

Supporting Information

References

1) Ali Y. M.; Volkan K.; Gizem K. Ö.;Abdullah B.; Nesrin H.; Mehmet E. S., Smartphone-based colorimetric detection via machine learning, *Analyst*, **2017**, *142*, 2434, doi: 10.1039/C7AN00741H.
2) Saptami, R.; Protik, C.B.; Md, A.H.; Md, R.I.; John, C. Polynomial regression of multiple sensing variables for high-performance smartphone colorimeter. *OSA Contin.* **2021**, *4*, 374–384.
3) Antela, K. U.; Sáez-Hernández, R.; Cervera, M. L.; Morales-Rubio, Á.; Luque, M. J. Development of an automated colorimeter controlled by Raspberry Pi4, *Analytical Methods*, **2023**, *15*, 512-518, doi: 10.1039/D2AY01532C.
4) Singh, H.; Singh, G.; Mahajan, D. K.; Kaur, N.; Singh, N., A Low-Cost Device for Rapid 'Color to Concentration' Quantification of Cyanide in Real Samples Using Paper-Based Sensing Chip, *Sens. Actuators, B*, **2020**, *322*, 128622.
5) (a) Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M., *Ore Geol. Rev.*, **2015**, *71*, 804–818; (b) Saini, C.; Mahato, K. D.; Azad, C.; Kumar, U., *Proc. Int. Conf. Artif. Intell. Appl. (ICAIA), ATCON-1*, **2023**.
6) (a) Lee, T., Lee, H.-T., Hong, J., Roh, S., Lee, K., Choi, Y., Hong, Y., Hwang, H.-J.; Lee, G., *Anal. Methods*, **2022**, *14*, 4749–4755.; (b) Kumar, A.; Mahato, K. D.; Azad, C.; Kumar, U.**,** *Proc. Int. Conf. Adv. Global Eng. Challenges (AGEC)*, **2023**, 118–123.
7) Jackman, P., Sun, D.-W. and ElMasry, G., *Meat Sci.*, 2012, *91*, 402–407.
8) Langsetmo, L.; Schousboe, J. T.; Taylor, B. C.; Cauley, J. A.; Fink, H. A.; Cawthon, P. M.; Kado, D. M.; Ensrud, K. E.**,** *JBMR Plus*, **2023**, *7*, e10757.
9) Mahato, K. D.; Das, S. S. G. K.; Azad, C.; Kumar, U.**,** *APL Mach. Learn.*, **2024**, *2*, 016101.
10) Pelligrini, A., Thomas, U., and Bramaz, N., *J. Appl. Microbiol.*, 1997, 82, 682–688