Colorimetric detection of an analyte with a smartphone

Under supervision of - Dr. Bharti Khungar

By - Nitish Bahl

For partial fulfillment of Lab Oriented Project

Name of Supervisor - Dr. Bharti Khungar, Asst. Professor, Dept. of Chemistry

Abstract - Colorimetry is the technique used to determine the concentration of colored compounds in a solution. The aim of the project includes developing machine learning models and algorithms, to predict colorimetric property such as pH of an analyte using primary colors(RGB) values. Using the ML model which predicts pH value, we can obtain real pH predicted value using a web application.

Keywords - pH, Colorimetric properties, RGB intensities, Machine Learning, KNN, SVM, Random Forest, web application, Flask, API development

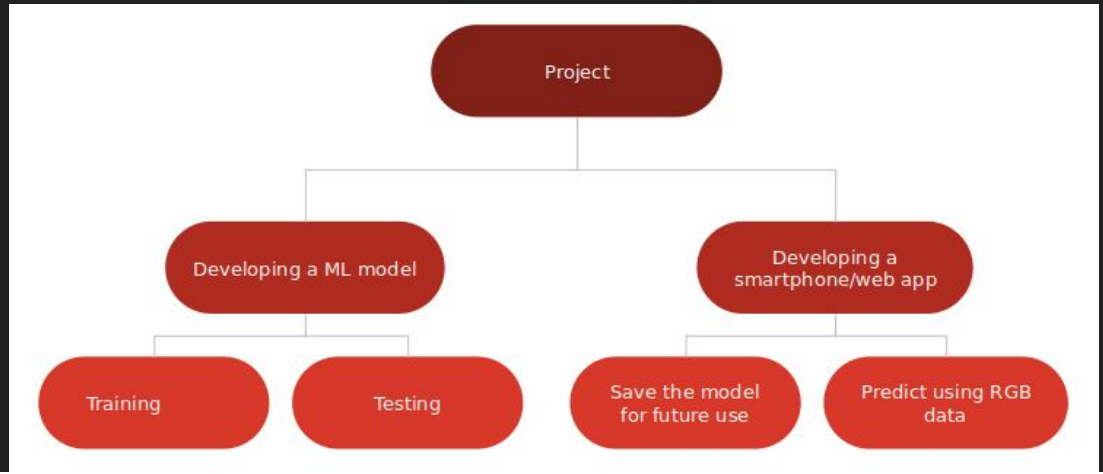
Determination of pH

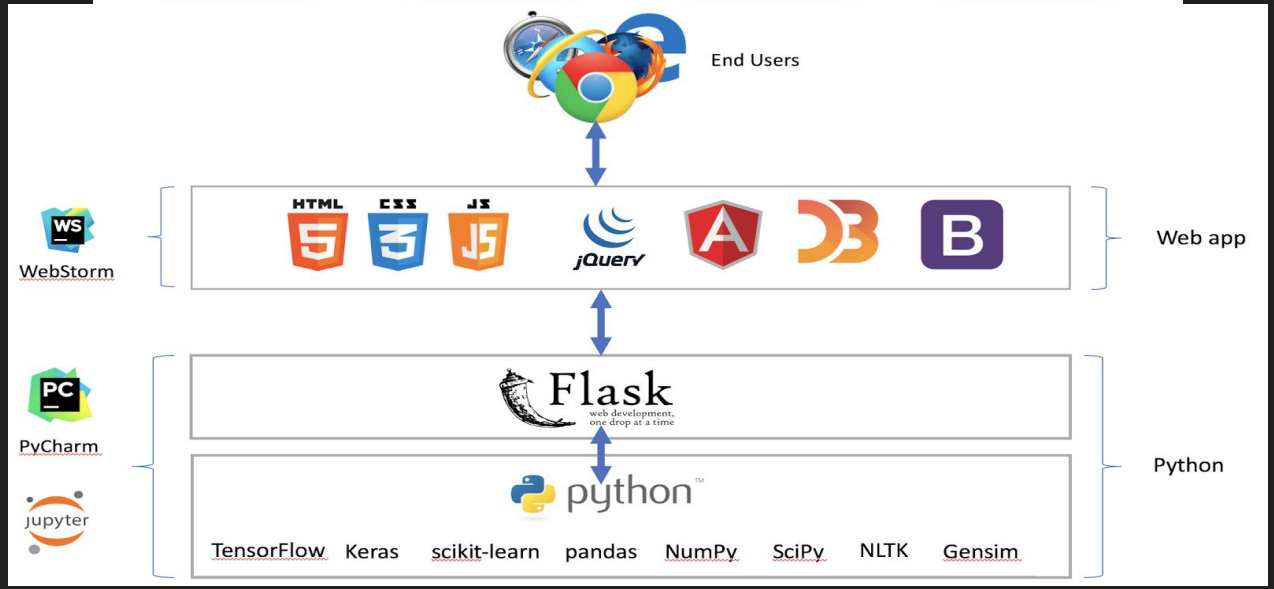
1. Indicator Methods (Colorimetric) - Indicators may be used to measure pH, by making use of the fact that their color changes with pH. Visual comparison of the color of a test solution with a standard color chart provides a means to measure pH accurate to the nearest whole number. More precise measurements are possible if the color is measured, using a colorimeter or spectrophotometer.

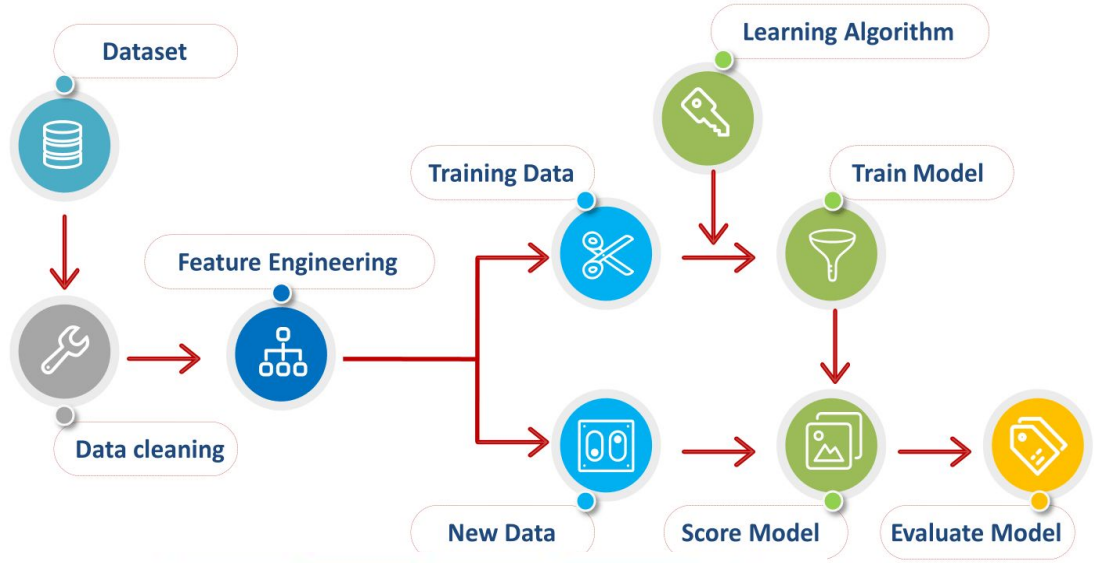
   Composition of Universal Indicator

Indicator	Low pH colour	Transition pH range	High pH colour
Thymol blue (first transition)	Red	1.2 – 2.8	Yellow
Methyl orange	Red	3.2 – 4.4	Yellow
Methyl red	Red	4.8 – 6.0	Yellow
Bromothymol blue	Yellow	6.0 – 7.6	Blue
Thymol blue (second transition)	Yellow	8.0 – 9.6	Blue
Phenolphthalein	Colourless	8.3 – 10.0	Fuchsia

2. Metal Electrode methods - A hydrogen electrode is made by adding platinum black to platinum wire or a platinum plate. It is immersed in the test solution and an electric charge is applied to the solution and the solution is saturated with hydrogen gas. The electrode potential is measured between the platinum black electrode and silver chloride electrode. This potential is inversely proportional to the pH of the solution.
3. Glass Electrode methods - The glass electrode method uses two electrodes, a glass electrode and a reference electrode, to determine the pH of a solution by measuring the voltage (potential) between them. This method is the one most commonly used for pH measurement since the potential quickly reaches equilibrium.


Project
Developing a ML model
Developing a smartphone/web app
Training
Testing
Save the model for future use
Predict using RGB data

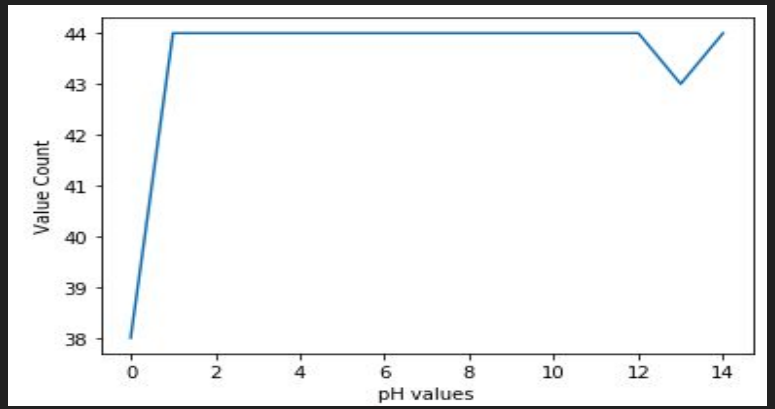

End Users
WebStorm
HTML
CSS
JS
jQuery
Web app
PyCharm
Flask
web development, one drop at a time
jupyter
python
TensorFlow
Keras
scikit-learn
pandas
NumPy
SciPy
NLTK
Gensim
Python




```
df.head()
```

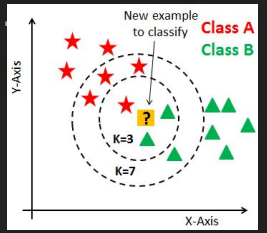
	blue	green	red	label
0	36	27	231	0
1	36	84	250	1
2	37	164	255	2
3	22	205	255	3
4	38	223	221	4

	blue	green	red	label
count	653.000000	653.000000	653.000000	653.000000
mean	89.290965	130.094946	120.655436	7.055130

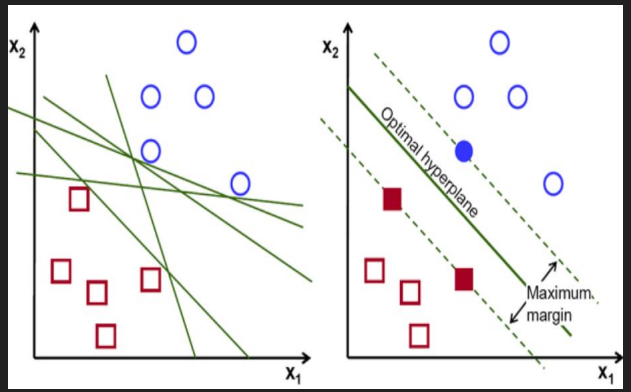


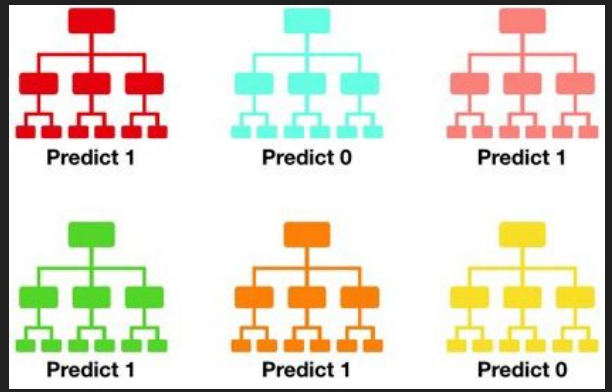
Training ML models

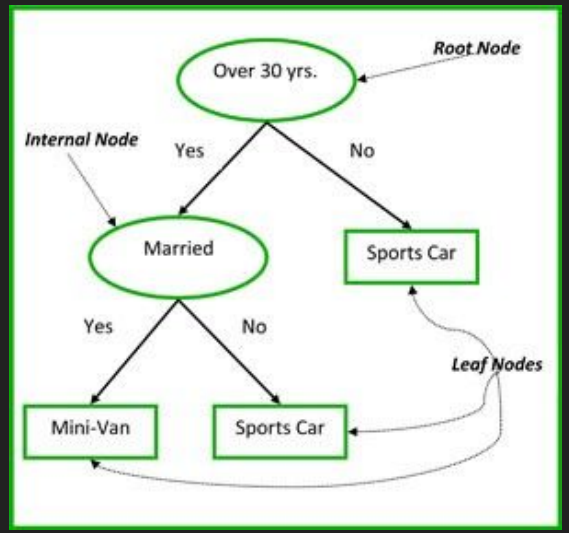
1. Linear Regression - Linear regression is used for finding linear relationship between target and one or more predictors. The core idea is to obtain a line that best fits the data.
2. Logistic Regression - Logistic Regression is used when the dependent variable(target) is categorical. Our case is Ordinal Logistic Regression.
3. K Nearest Neighbours - This algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with calculating the distance between points on a graph.



4. Decision Trees - Tree-like graph or model of decisions and their possible consequences, including chance event outcomes.It is a flowchart-like structure in which each internal node represents a “test” on an attribute.
5. Random Forest - Consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.
6. Support Vector Machines



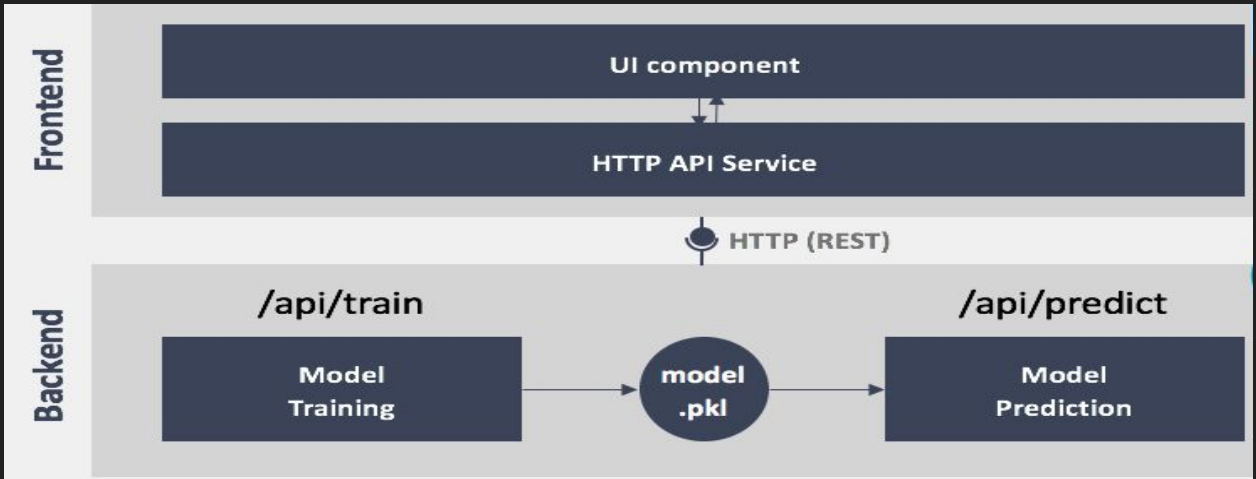


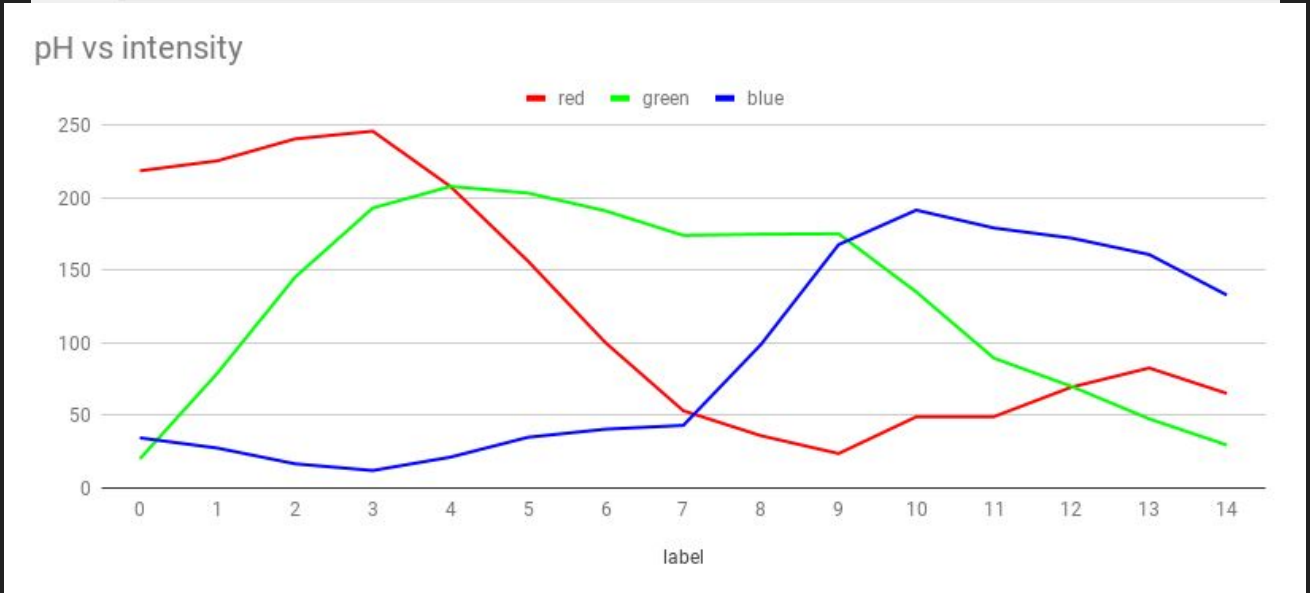


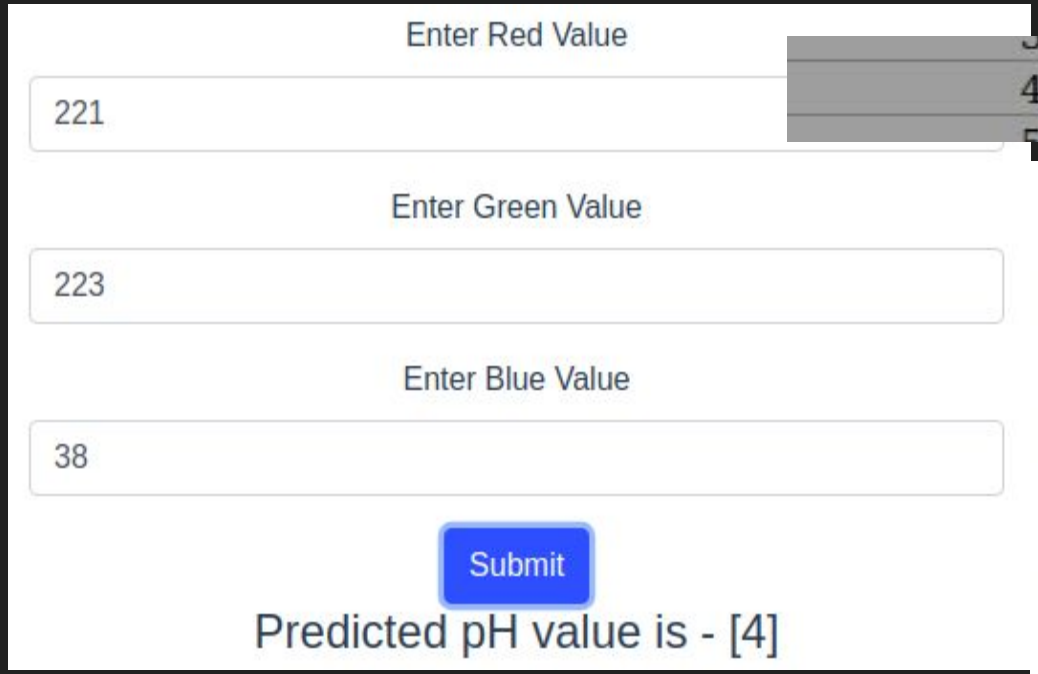
Brief summary of results of different algorithms

Optimal Algorithm	Accuracy(%)	Mean Error
Linear Regression	45	1.88
Logistic Regression	52	0.77
KNN	68	0.56
SVM	68	0.54
Decision Trees	66	0.54
Random Forest Classifier	72	0.51

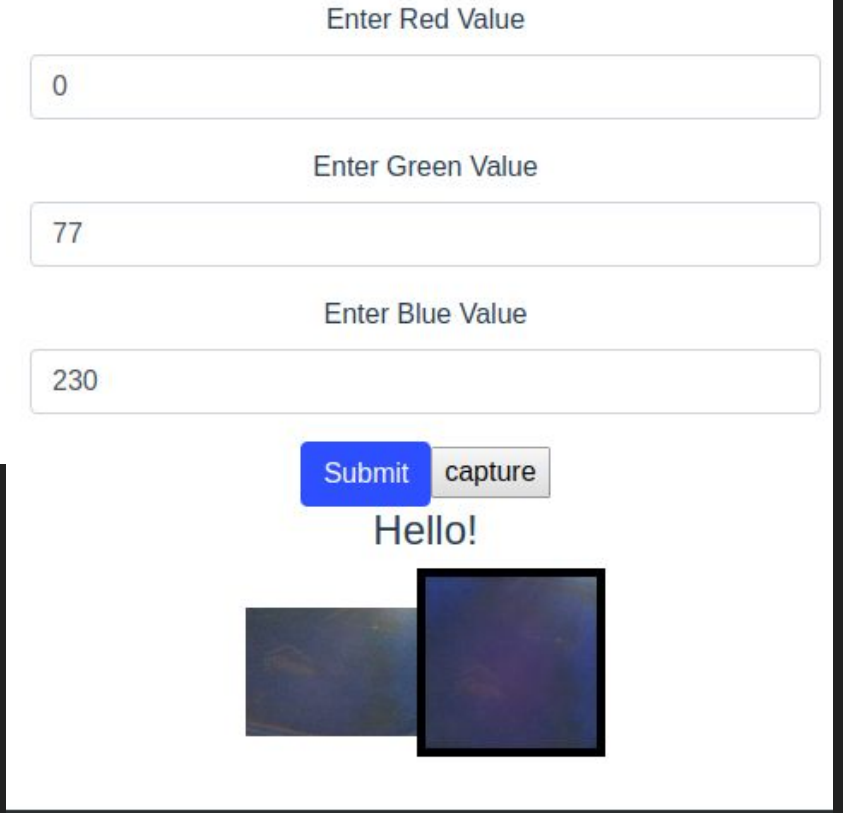
It can be seen that both accuracy and mean absolute error have stagnated once we shift to complex algorithms. Still, Random Forest Classifier has achieved the highest accuracy of 72% and just 0.51 mean error.


Frontend
UI component
HTTP API Service
HTTP (REST)
Backend
/api/train
/api/predict
Model Training
model .pkl
Model Prediction


pH vs intensity
red
green
blue
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
label



4	221	223	38

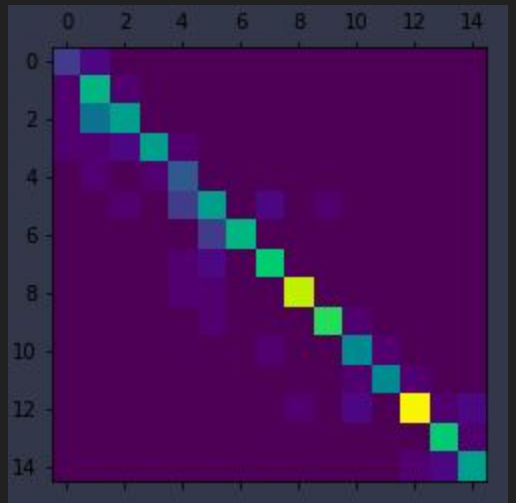


Actual Images of the web application.

1. The top image shows correct prediction, based on reference data.
2. The image on right shows application connected with camera to view live feed as well as captured image, which auto populates the RGB values.

Conclusion

By plotting a graph between Color Intensity and pH, we can correlate it physically that lower ranges of pH(0-5) have a high value of red intensity and low value of blue intensity and the opposite happens in case of pH range(7-14). Green intensity first increases and then decreases gradually.



References

1. Point-of-care colorimetric detection with a smartphone - Li Shen, Joshua A. Hagen, Ian Papautsky, September 2012
2. Machine Learning Algorithms, https://towardsdatascience.com/
3. Ph-recognition, recognize pH value based on image color - https://www.kaggle.com/robjan/ph-recognition
4. Errors in pH measurement with colorimetric indicators in low alkalinity waters - Terry A. Haines John J. Akielaszek Stephen A. Norton Ronald B. Davis, 1983