REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ACCURATE PH DETERMINATION BY IMAGE PROCESSING USING CONVENTIONAL PH- TEST STRIP

Farah EID

MASTER OF SCIENCE THESIS

Department of Chemical Engineering

Program of Chemical Engineering

Supervisor

Prof. Dr. Hasan SADIKOĞLU

November, 2024

REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ACCURATE PH DETERMINATION BY IMAGE PROCESSING USING CONVENTIONAL PH- TEST STRIP

A thesis submitted by Farah EID in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 08.11.2024 in the Department of Chemical Engineering, Chemical Engineering Program.

Prof. Dr. Hasan SADIKOĞLU
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Prof. Dr. Hasan SADIKOĞLU, Supervisor
Yildiz Technical University _____

Prof. Dr. Nil ACARALI, Member
Yildiz Technical University _____

Assist. Prof. Dr. Özgün YÜCEL, Member
Gebze Technical University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study, and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of "Accurate Ph Determination By Image Processing Using Conventional Ph- Test Strip" supervised by my supervisor, Prof. Dr. Hasan SADIKOĞLU. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Farah EID

Signature

Dedicated to my family and friends

ACKNOWLEDGEMENTS

I would like to express my gratitude and thanks to Prof. Dr. Hasan SADIKOĞLU and to Res. Asst. Mert Akın İNSEL for their help during every step of my thesis study. I am grateful to all of the professors and instructors at Yıldız Technical University's Chemical Engineering Department for helping me to develop my knowledge and abilities in this field. I want also to express my gratitude to my beloved family, who have supported me during every stage of my life and during my academic career. I am grateful to my supporting husband Ömer ARAR and my beautiful angel Khalid ARAR for serving as my inspiration source. Also, my thanks are for my best friends how were motivating me all the time.

Farah EID

TABLE OF CONTENTS

LIST OF SYMBOLS ix
LIST OF ABBREVIATIONS x
LIST OF FIGURES xi
LIST OF TABLES xiii
ABSTRACT xiv
ÖZET xvi
1 . INTRODUCTION 1
1.1 Literature Review 1
1.2 Objectives of The Thesis 5
1.3 Original Contribution 5
2 . THEORY 6
2.1 pH Scale 6
2.2 pH Measurement Methods 7
2.2.1 Electrochemical methods 7
2.2.2 Optical methods 7
2.3 pH Paper 8
2.3.1 Litmus Paper 8
2.3.2 Universal Test Strips 8
2.4 Image Processing 9
2.4.1 Image Processing Components 9
2.4.2 Image Processing Fundamental Steps 9
2.5 Machine Learning 10
2.6 Machine Learning Algorithm Types 11
2.6.1 Supervised Learning 12
2.6.2 Unsupervised Machine Learning 13
2.6.3 Reinforcement learning 14
3 . METHODOLOGY 15
3.1 Data Collection 15

3.1.1 Solutions Preparing 15
3.2 pH Strips Immerging 21
3.3 Photos Capturing 22
3.3.1 At laboratory ambient light 22
3.3.2 Using special prepared black box 22
3.4 MATLAB Coding 23
3.5 Model Selection 23
3.5.1 Linear Regression 24
3.5.2 Decision tree 24
3.5.3 Support vector machine 25
3.5.4 Kernel 26
3.5.5 Gaussian Process Regression 27
3.5.6 Ensemble 27
3.5.7 Neural Network 29
3.6 Evaluation Method 30
3.6.1 Cross-validation 30
3.6.2 Performance Metrics 31
3.7 Model Optimization 33
3.7.1 Bayesian Optimization 33
4 . RESULTS AND DISCUSSION **35**
4.1 pH Solutions and pH Strips Results 35
4.2 MATLAB GRB Results 37
4.3 Regression Learner 38
4.3.1 Chosen machine learning models 38
4.4 Regression Results 39
4.4.1 Reference paper inside the black box 39
4.4.2 Reference paper in ambient laboratory light 43
4.4.3 Experiment Set Inside the Black Box Results 43
4.4.4 Experiment Set in Ambient Laboratory Light Results 45
4.5 Regressions Findings 48
4.6 Validation with the test experiment set 48
4.7 Comparison with other study 49
4.8 Future Aspects 50

5 . CONCLUSION **51**
REFERENCES **52**
PUBLICATIONS FROM THE THESIS **58**

LIST OF SYMBOLS

R^2	Coefficient of Correlation
M	Molar
pH	Power of Hydrogen Ion
V	Volume in Liter

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
BO	Bayesian Optimization
CIE	Commission International de l'Éclairage
DT	Decision Tree
GRB	Green Red Blue Colors Value
GPR	Gaussian Process Regression
HPO	Hyperparameter Optimization
KNN	K-Nearest Neighborhood
MAE	Mean Absolute Error
MSE	Mean Square Error
ML	Machine Learning
NN	Neural Network
PC	Personal Computer
RMSE	Root Mean Square Error

LIST OF FIGURES

Figure 2.1 pH scale 6
Figure 2.2 Experimental methods to measure pH values 7
Figure 2.3 Basic element of digital image processing system 9
Figure 2.4 Machine learning types 11
Figure 2.5 Graphically comparison between Regression and Classification 12
Figure 3.1 Main 4 solutions prepared. 15
Figure 3.2 Chosen pH strips 21
Figure 3.3 Samples of solutions prepared. 21
Figure 3.4 Photos capturing set at laboratory ambient light. 22
Figure 3.5 Special prepared black box. 22
Figure 3.6 MATLAB code of pixels selection 23
Figure 3.7 Decision tree structure 25
Figure 3.8 Schematic diagram of a support vector machine (SVM) 26
Figure 3.9 Illustration of data projection of a kernel function 26
Figure 3.10 Bagging (a) and boosting (b) ensemble methods 28
Figure 3.11 Neural Network structure 29
Figure 3.12 K-Fold cross validation 31
Figure 3.13 General flowchart of Bayesian optimization 34
Figure 4.1 Used pH meter. 35
Figure 4.2 pH strips tested with (0, 1, 2, 3, 4, 5, 6, 7) pH degree solutions 36
Figure 4.3 pH strips tested with (7.5, 8.5, 9.5, 10.5, 11.5, 12.5, 13.5) pH degree solutions. 36
Figure 4.4 MATLAB regression learner. 38
Figure 4.5 Selected regression models used. 39
Figure 4.6 GPR plot for validation predicted vs. actual with 232 pixels RGB values feed. 40
Figure 4.7 GPR plot for validation predicted vs. actual with 696 pixels RGB values feed. 42

Figure 4.8 GPR plot for validation predicted vs. actual with 1392 pixels RGB values feed. 44

Figure 4.9 GPR plot for validation predicted vs. actual with 1392 pixels RGB values feed. 46

Figure 4.10 NN plot for validation predicted vs. actual with 1392 pixels RGB values feed. 46

Figure 4.11 Test predicted vs actual predictions. 47

Figure 4.12 Plot of validation results for predicted vs. actual. 49

LIST OF TABLES

Table 1.1 Summary of relevant studies.....2

Table 1.2 Summary of relevant studies (continued). 3

Table 1.3 Summary of relevant studies (continued). 4

Table 3.1 (0, 1, 2, 3, 4, 5, 6) pH Solutions preparing amounts of HCl and water. 17

Table 3.2 (0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5) pH Solutions preparing amounts of HCl and water.....18

Table 3.3 (14, 13, 12, 11, 10, 9, 8) pH Solutions preparing amounts of NaOH and water.....19

Table 3.4 (13.5, 12.5, 11.5, 10.5, 9.5, 8.5, 7.5) pH Solutions preparing amounts of NaOH and water 20

Table 4.1 Example of RGB data table constructed. 37

Table 4.2 Machine learning examined models. 38

Table 4.3 Validation evaluation of reference paper inside the black box with 232 pixels RGB values feed.....40

Table 4.4 Validation evaluation of reference paper inside the black box with 464 pixels RGB values feed.....41

Table 4.5 Validation evaluation of reference paper inside the black box with 696 pixels RGB values feed.....41

Table 4.6 Validation evaluation of reference paper inside the black box with 696 pixels RGB values feed (continued). 42

Table 4.7 Validation evaluation of reference paper in ambinet laboratory light with 696 pixels RGB values feed.....43

Table 4.8 Validation evaluation metrics of reference paper inside the black box with 1392 pixels RGB values feed.....44

Table 4.9 Validation evaluation metrics of reference paper in ambient laboratory light with 1392 pixels RGB values feed. 45

Table 4.10 Test set results. 47

Table 4.11 Optimzation results. 48

Table 4.12 Test set performance metrics. 49

Table 4.13 Compression between our study and relevant study from literature...50

ABSTRACT

Accurate Ph Determination by Image Processing Using Conventional Ph- Test Strip

Farah EID

Department of Chemical Engineering

Chemical Engineering Program

Master of Science Thesis

Supervisor: Prof. Dr. Hasan SADIKOĞLU

pH measurement is considered to be an important measurement with a variety of application in many areas. It is essentially determining if the aqueous solutions are alkaline or acidic. Using pH strips is the most common and traditional way for measuring pH. Although they are less accurate than pH meters, they are known as a cost-effective and affordable measurement method that non-experts can use. The main disadvantage of using pH strips is the high possibility of having scientific error due to the limited precision and subjectivity to human judgment when evaluating the value of strip. Nonetheless, having access to modern technology and cell phones provides a significant benefit in overcoming this obstacle. Many studies

use image processing to measure pH by examining the pH strip's color using different techniques such as GRB value or CIE color space. In this study, machine learning algorithms are utilized in order to construct a regression model for determining pH value by analyzing the pixel RGB color data of pH strips using MATLAB software. Several machine learning algorithms (such as linear regression, Gaussian process regression, Nurel network, Ensemble, Support vector machine and decision tree) were utilized to find the most appropriate model for our data. In this context, 29 different pH solutions were prepared by dilution method, 2 experiment sets under different light conditions were used to capture the pH strips photos, MATLAB software used regression data as RGB values, and the regression is done also by the use of MATLAB. Some evaluation metrics such as RMSE, MAE, MSE and R^2 were calculated to figure out the most appropriate model. The best prediction was given by the GPR model with R^2 equal to 0.9754392. The suggested model uses 10-fold cross-regression validation method and Bayesian optimization. The obtained model is examined with a new set of data and proved its efficiency. The study's findings suggest that machine learning algorithms are effective tools for producing reliable models for estimating pH value.

Keywords: pH strips, Image Processing, Cross-regression validation, Machine learning.

ÖZET

Geleneksel Ph Test Kağidi Kullanilarak Görüntü İşlem İle Doğru Ph Belirleme

Farah EID

Kimya Mühendisliği Anabilim Dalı

Kimya Mühendisliği Programı

Yüksek Lisan Tezi

Danışman: Prof. Dr. Hasan SADIKOĞLU

PH'ın belirlenmesi, çeşitli uygulamalarda kullanılan kritik bir ölçüm olarak kabul edilir. Esas olarak sulu çözeltilerin asitliğini değerlendirmek ve bunları asidik veya alkali olarak sınıflandırmak için kullanılır. PH ölçümü için çeşitli yöntemler vardır. En geleneksel ve yaygın olarak kullanılan yöntem pH şeritlerinin kullanılmasıdır. Doğruluğu pH metrelere göre daha az olsa da, uzman olmayanların da kullanabileceği ve uygun fiyatlı bir ölçüm yöntemi olarak biliniyorlar. pH şeritleri kullanmanın ana dezavantajı, şerit değerini okurken sınırlı hassasiyet ve insan yargısının öznelliği nedeniyle bilimsel hata olasılığının yüksek olmasıdır. Ancak günümüzde ileri teknolojinin ve cep telefonlarının varlığı bu zorluğun üstesinden gelmek için büyük bir avantaj sunmaktadır. Birçok araştırmacı, GRB değeri veya CIE renk uzayı gibi çeşitli yöntemler kullanılarak pH şeridinin renginin görüntü işlenmesiyle pH değerinin belirlenmesinin mümkün olduğunu bulmuşlardır. Bu

çalışmada, pH şeritlerinin piksel rengini MATLAB yazılımı kullanarak analiz ederek pH değerini belirlemek için bir regresyon modeli oluşturulmuştur. Verilerimiz için en uygun modeli bulmak için çeşitli makine öğrenme algoritmaları (liner regresyon, Gaussian process regression, Nurel network, Ensemble, Support vector machine and decision tree gibi) kullanıldı. Bu bağlamda, seyreltme yöntemi ile 29 farklı pH çözeltisi hazırlanmış, pH şeritlerinin fotoğraflarını çekmek için farklı ışık koşulları altında 2 deney seti kullanılmış, MATLAB yazılımı regresyon verilerini RGB değerleri olarak kullanmış ve regresyon da MATLAB kullanılarak yapılmıştır. En uygun modeli bulmak için RMSE, MAE, MSE ve R^2 gibi bazı değerlendirme metrikleri hesaplanmıştır. En iyi tahmini, R^2'si 0,9754392'ye eşit olan GPR modeli vermiştir. Önerilen model, 10 katlı çapraz regresyon doğrulama yöntemini ve Bayes optimizasyonunu kullanmaktadır. Elde edilen model yeni bir veri seti ile incelenmiş ve etkinliği kanıtlanmıştır. Sonuçlar, görüntü işleme ve makine öğrenmesi algoritmalarının tahmin alanında oldukça etkili olduğunu göstermektedir.

Anahtar kelimeler: pH şeritleri, Görüntü İşleme, Makine öğrenme, Çapraz regresyon doğrulama

1
INTRODUCTION

1.1 Literature Review

pH value is a critical parameter in many areas including chemistry, food manufacturing, agriculture, health etc., since it is the measure for acidity and alkalinity of aqueous solution. Traditionally, pH meters, pH strips and indicator solutions have been employed in pH value measuring [1]. Determining the pH value using pH strips is a simple, economical and effective method that can be used by farmers, aquarists and people who want to take care of their health by themselves [2], [3]. Despite the advantages of using pH strips, there are some drawbacks that affect their accuracy and reliability such as limited precision and subjectivity to human judgments. However, the presence of image processing enables us to overcome these drawbacks and enhance the reading of pH strips by using automatic algorithms that give the most precise reading [4]. These algorithms can analyze the images of pH strips taken by mobile phone and extract a database depending on color differentiation, RGB value or color wavelength to use in the determination of pH value [1], [2]. The pH reading system mainly consists of capturing the pH image and data analysis software that offer a reliable and comprehensive pH value measurement [5]. By image processing technique the colorimetric data can be interpreted, and the relevant color data can be extracted which allows us to make correlation to the corresponding pH value [6]. The purpose of this study is to investigate machine learning algorithms in order to choose the appropriate algorithm that predict accurate pH value. This will be done by analyzing captured images of pH strips using a iPhone XR camera, MATLAB software, color analysis of pixels, and machine learning algorithms with a 10-fold cross validation.

In literature there are different researches that have revealed the feasibility and efficiency of using image processing for pH value determination. The following table provides an overview of the most relevant studies which were done in the field of pH determination using image processing.

Table 1.1 Summary of relevant studies

Study Name/ Year/ Reference	Summery	Methodology	Limitations
Classification of pH Levels using a Mobile Phone /2009/[6]	The study suggests an automatic system for estimating pH value by taking a picture of the tested pH strips beside with the pH scale reference using mobile Camera and then doing image processing of this picture to determine the pH value of the strip.	9 solutions of different pH were prepared in laboratory, the pH of these solution was tested using a pH test strips then 3 Step including software and algorithms were done to determine pH value: first by detecting of the color target by Sobel edge detection, second applying color transform using pixel from RGB color values and third determine the pH level.	Despite of that the estimated pH values were close to true, but it is does not give enough precise value. also the results can be changed by changing the size of the photos. The study was performed in 2009 and mobile phone camera resolution at that time was not good enough.
Evaluation of pH Value by Using Image Processing /(2017)/ [7]	The study explains a titration experiment of HCl with NaOH, color changes of solution were observed and analyzed by image processing, finding a linear relationship between increasing red values and pH value, revealing the potential use of image processing for pH determining.	Titration of HCl solution with NaOH solution was performed in the presence of an indicator, Images of the solution at different titration steps were taken and analyzed by software, measuring the average RGB values and finding the most convenient color that has a linear correlation with pH value, pH and pOH Equations were used.	Thers is no obvious limitations mentioned in this study but there is a mention that image processing approach needs further validation and testing before relying on to be a reliable alternative way to traditional pH measurement.

Table 1.2 Summary of relevant studies (continued).

Smartphone-Based Automatic Measurement Method for Colorimetric pH Detection Using a Color Adaptation Algorithm/ 2017/ [8]	The research created a colorimetric pH detection system for smartphones that adjusts for ambient light and image sensor characteristics using a 3D printed miniature light box and color calibration. For high-resolution pH estimate, the system employs non-rigid u'v' curve interpolation and a third-order polynomial model. Additionally, a technique for automatically obtaining color information is devised.	By using a 3D-printed small light box to isolate ambient light, photos were taken of pH strips immersed in buffer solutions with a printed comparison chart from 3 Android cell phones. Using the C++ programming language, an Android application was developed to take pictures and measure the pH. To match the captured color with the reference color in the CIE 1976 u'v' color space, a color conversion module was driven. The pH was then calculated using non-rigid u'v' curve interpolation, which produced high-resolution pH estimates, and an automatic contour extraction-based color recovery algorithm was put into place.	The proposed system can give acceptable results but still there is some limitation summarized in the used paper-printed comparison differs in its nature from the real pH strips immersed in solution, Additionally, there isn't a precise, set place to take the pictures, which might cause variations in the outcome. The performance of color adaptation can be weakened by saturated color information from automated brightness control and white balancing. Also, the resolution of smart phones camera has been improved a lot since 2017.

Table 1.3 Summary of relevant studies (continued).

Quantitative pH Determination Based on the Dominant Wavelength Analysis of Commercial Test Strips / 2021/ [1]	The study developed a smart-phone application that can determine pH value with comparatively high accuracy by analyzing the captured images of pH strips for the dominant and complementary wavelength of the color which is a promising tool to be used and developed in future.	Preparation of standard buffer solutions with known pH value was done, 3D-printed optical attachment with a surface light source and macro lens was employed to capture the pH strip's images, by the use of Java language development of android App that make a correlation between the dominant wavelength of the color in captured image and the pH value was done, eventually the effectiveness of the developed App was examined by the use of pH meter and real-world samples.	Even though the study gives very acceptable results and done in a really detailed and professional manner, the 10-14 pH range was not included in the solutions prepared and that is an important part to be include,
Mobile App to Quantify pH Strips and Monitor Titrations: Smartphone-Aided Chemical Education and Classroom Demonstrations /2023/ [9]	The study built on the work of the previous study [1], describing the development and testing of a mobile app for high school and undergraduate students to determine solution pH using pH strips or perform acid-base titrations.	Building on the work of the previous study, the development of 2 variants of a mobile phone app was done, one is analyzing photos of pH paper strips, while the other analyzes photos of solutions with a universal pH indicator. Calibration of the App was done by using standard buffer solutions.	There was a limitation in calibration of both variants of the app due to the absence of some ranges of pH; 0−2, 4.5−6, 8−9 and 12−14 (for short-range pH paper) and 0−2 and 13−14 (for universal pH paper) for paper reader and 0 to 4 and 10−14 for indicator reader

1.2 Objectives of The Thesis

The aim of this study is to utilize machine learning algorithms to build a regression model to determine pH value of an examined solution from the captured image of the pH test strip dipped into. By using colorimetric analysis of manually selected pixels from pH strips photos and by the use of MATLAB regression learner with machine learning algorithms such as (linear regression, Gaussian process regression, Nurel network, Ensemble, Support vector machine and decision tree).

1.3 Original Contribution

The original contribuation from this study is the whole methodology followed, used the most optimized factors that affecting the model performance such as choosing the most proved efficient pH test strips type with 4 color rectangular besides performing experiments in 2 different light circumferences also using iPhone Camera with its known high quality and high color resolution. Also, integration of multiple ML algorithms does not just rely on a single linear method. Besides using enhanced model optimization techniques that maximize model performance and minimize prediction error demonstrating that even with minimal equipment and standardized conditions, precise pH measurement can be achieved.

2
THEORY

2.1 pH Scale

The concept of the pH scale, which was formulated more than a century ago by the Danish chemist Søren Sørensen, has become indispensable in many areas of life today [10]. It is not only of fundamental importance in chemistry, agriculture and environmental sciences, but also in product manufacturing, such as medicines, food, paintings, etc.[11]. pH as a term comes from Latin and stand for (pondus hydrogenii) which means weight of hydrogen [12]. The pH scale is basically a measure of the acidity or alkalinity of a solution depending on the concentration of hydrogen ions. It is a logarithmic scale ranging from 0 to 14, with 0 belonging to the strong acids, 14 to the strong bases, and 7 to the neutral solution [13].

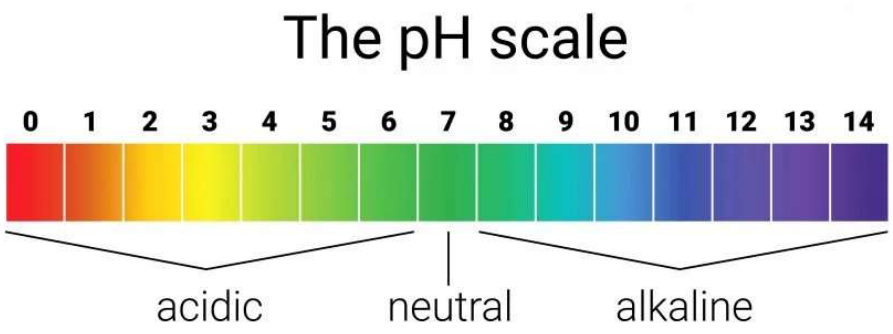


Figure 2.1 pH scale [14]

Basically, it is described by the negative logarithm of hydrogen ion concentration. It is expressed as an equation (2.1)

$$pH = -\log [H\text{^}+\,] \tag{1.1}$$

where $[H^+]$ represents solvated protons concentration [14]. Chemists realized that measuring hydrogen ion concentration was actually measuring hydrogen ion activity [15].

2.2 pH Measurement Methods

Various methods are utilized in measuring pH value, including electrochemical and non-electrochemical methods [15].

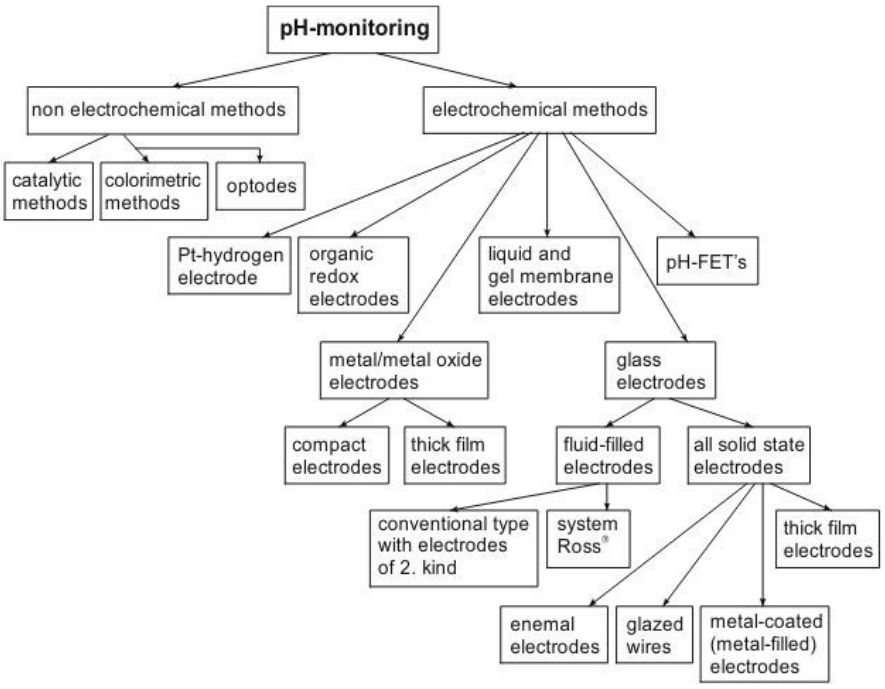


Figure 2.2 Experimental methods to measure pH values [17].

Among these methods, pH meter as an electrochemical and pH strips as colorimetric method are the simplest and commonly used methods.

2.2.1 Electrochemical methods

Reactions including [H^+] or [OH^-] ions are the most used in a lot of electrochemical methods for determining pH values, like hydrogen electrodes, organic redox electrodes, metal/metal oxide electrodes and glass electrodes, The Nernst equation serves as the foundation for these electrodes' operation. [16].

2.2.2 Optical methods

Optical methods have advantages over electrochemical methods for pH measurement, using organic dye molecules for acid-base titrations, indicator papers, and fiber-optic probes for localized pH within living cells [15].

Among the previous methods, indicator papers or pH strips are considered to be the simplest and most cost-effective method for measuring the pH value [17].

2.3 pH Paper

pH papers are a straightforward, efficient, and cost-effective method for assessing pH levels.[18]. The decade of 1930s is the first time the pH indicator paper was developed by impregnating the strip in a mixture of dyes that change color with pH degree changing but it was difficult to read and was used only on a narrow scale, until 1960 were it became more precise after the synthesis of new dyes, between the 1970s and 1980s, pH papers evolved with new formulations and coatings, including gel or polymer coated ones to prevent dye bleeding or running [19]. Its principle of working is too simple, they consist of a strip of paper coated with indicator dyes chosen for the pH range. The pH is determined by dipping the strips in the solution then comparing the strip's color to a reference's calibration chart, accuracy of pH degree is up to 1 or 0.5 units depending on the strip's type [15]. pH paper types are related to substances that aimed to be examined also to the range of pH, the most common types of pH paper are as follows:

2.3.1 Litmus Paper

Litmus paper is one of the most popular methods of pH determining. As its name it is made of litmus dye**.** There are two types of litmus paper: blue and red paper. Where the blue one becomes red when an acidic medium, and the red one becomes blue whether the substance is basic or alkaline [20]. An important thing to mention is that litmus paper does not determine the specific pH degree it is just about determining a substance whether to be acidic or alkaline whereas litmus pH test strips dose [21].

2.3.2 Universal Test Strips

They are generally used for a wide range of substances and offer a kind of precise pH determining ranging from 0 to 14, with sometimes 0.5 pH degree interval [22],[21]. There are also narrow range pH strips that only measure specific pH range like from 6 to 9 [23].

2.4 Image Processing

Image processing is a technique that enhances the quality and extracts useful information from images, treating them as two-dimensional signals where Inputs include photographs or videos, and the output potentially improved images or associated features or information extracted form images [24]. often using advanced mathematical concepts and gray level arrays as input [25]. It is crucial in various applications like face recognition, object detection, and image compression [26]. There are several ways to process a picture, including object segmentation [27], image arithmetic, image enhancement, and study of grey values [28].

2.4.1 Image Processing Components

To perform image processing several components should be used, essentially an acquisition device to have the image, processing unit; computer and software to treat the image, display device also sometimes it necessary to use hardcopy equipment and networking [29], [30].

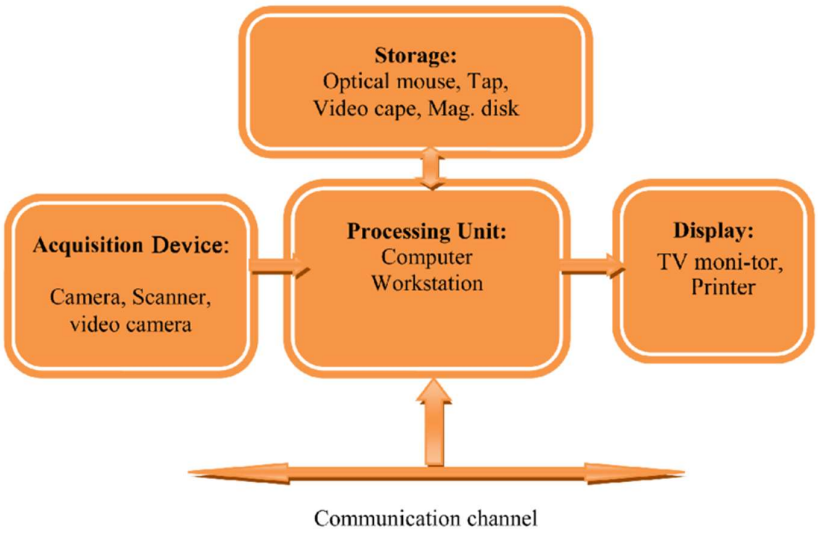


Figure 2.3 Basic element of digital image processing system [29].

2.4.2 Image Processing Fundamental Steps

Digital image processing entails several processes to modify and analyse pictures for a variety of uses, including pattern detection, segmentation, and enhancement, [31]. The fundamental steps in image processing are summarized as follows:

- Image Acquisition.
- Image Enhancement.
- Image Restoration.
- Color Image Processing.
- Wavelets and Multi-Resolution Processing.
- Image Compression.
- Morphological Processing.
- Image Segmentation.
- Representation and Description.
- Object recognition.
- Knowledge Base [31].

Image processing techniques significantly impact technologic organizations by offering various benefits such as digital image availability in various formats, improved human interpretation, machine interpretation information extraction, pixel manipulation, easy storage and retrieval, and easy electronic transmission to third-party providers [30], In our study we are going to utilize color image processing to determine pH value from pH strips images.

2.5 Machine Learning

Machine learning is a branch in artificial intelligence, specifically is a scientific study of algorithms and statistical models used in computer systems for doing tasks like data mining, image processing, and predictive analytics, enabling automatic work without programming once the computer learned from previous data [32]. In other words, the goal of computer science's machine learning discipline is to make computers capable of learning without explicit programming [33]. By using different algorithms to discover the mathematical relationship between set of data [34]. Machine learning involves several steps:

- Data gathering.
- Data pre-processing.
- Choose the model.

- Train the model.
- Test model.
- Optimize the module.
- Prediction [35].

There are a lot of applications of machine learning in life sectors such as healthcare, finance and almost every sector [36].

2.6 Machine Learning Algorithm Types

Machine learning algorithms are essentially categorized into 3 types depending on data availability which is typically regarded as the most important factor in building a machine learning model [37] [38]. The main types of machine learning are as follows:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Each type contains a number of algorithms as summarized in Figure 4.

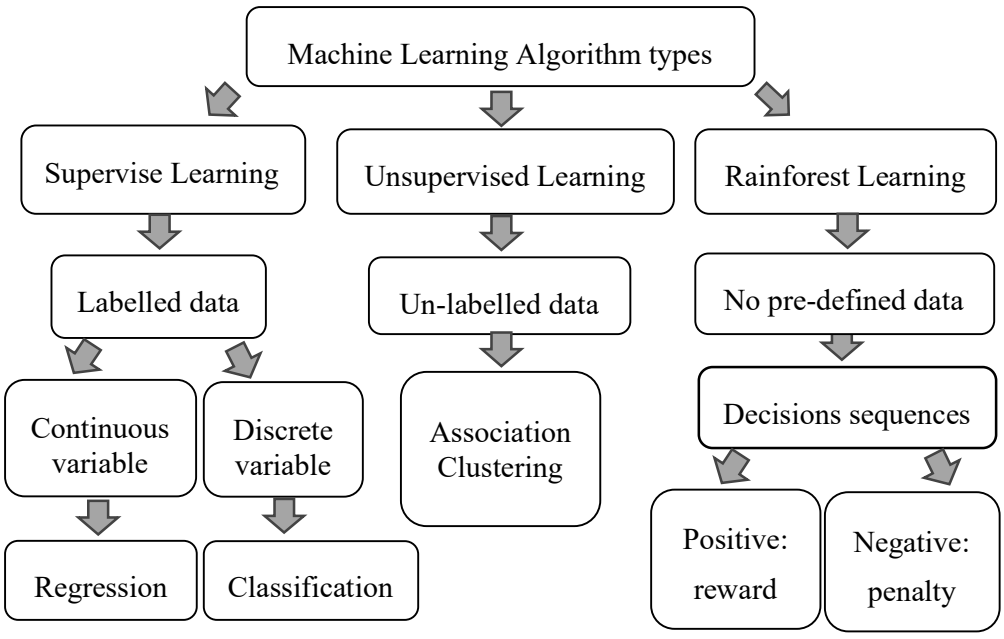


Figure 2.4 Machine learning types.

In this study supervise learning types of machine learning are going to be used since the data (input and output) are known and given to the algorithm.

2.6.1 Supervised Learning

As its name supervised learning suppose the presence of a supervisor that can give guidance and feedback, this supervisor is the training data [38], Supervised learning deal with labeled data; input and output are known and labeled so it can predict results [39] it takes place when certain objectives are determined to be achieved from a given set of inputs, The model generated can be utilized mainly for 2 tasks: prediction of future input data responses and inference to understand the relationship between inputs and output [37], as summarized in figure 4 according to the problem's variable to be continuous or discrete supervised learning tasks divided into 2 category [40]:

- Classification
- Regression

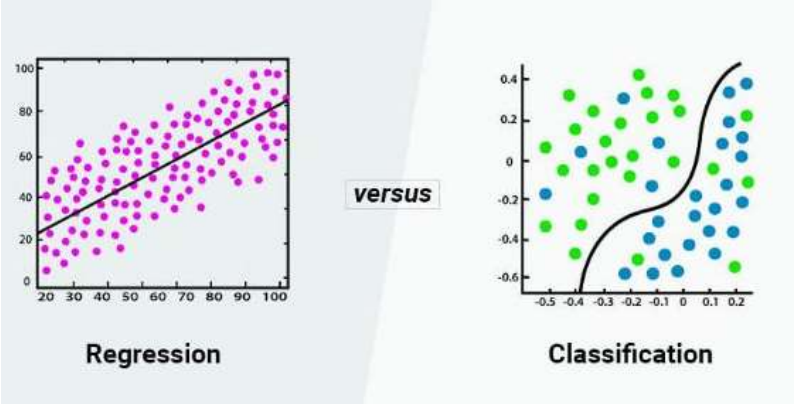


Figure 2.5 Graphically comparison between Regression and Classification [41].

A supervised learning algorithm constructs a predictive model, which, after training, predicts the most probable labels for new samples in the testing set, based on the target's nature [40].

2.6.1.1 Regression

Regression is a supervised machine learning technique that takes place in predicting continuous variables by using statistical methods that discover correlations between dependent and independent variables and choosing a mapping function to make the

input and output variables match, it used in weather state and real estate prices [41]. The following are the most common types of machine learning algorithms that are used for regression:

- Linear.
- Polynomial.
- Decision Tree.
- Random Forest.
- Support Vector Regression.

2.6.1.2 Classification

Classification is another supervised machine learning technique that takes place in predicting a discrete value, the algorithm functions try to predict the correct group or category of the input data and match each input to its output category. it deals with binary values such as yes or no 1 or 0. In other words; classification algorithms forecast the likelihood that an event will occur.by fitting data to a logit function [42], The following are the most common types of machine learning algorithms that used for regression:

- Decision Tree Classification.
- K-Nearest Neighbors.
- Random Forest Classification.
- Support Vector Machines.
- Naïve Bayes.
- Logistic Regression.

Although the last type of name is regression, but it should not be confused with the regression algorithms because it used to determine whether it is probable that the output variable's realization matches to the appropriate category [43].

2.6.2 Unsupervised Machine Learning

Un-supervised learning deals with unlabeled data without guidance which make it more complicated and challenging than supervised type, the data in this type is only input and the machine has to discover the hidden pattern and extract the rules from it [44], It is helpful for classifying data, finding underlying patterns, and recognizing

abnormalities, especially for situations when labelled data is hard to access or unavailable. As it mentioned earlier in figure 4 unsupervised learning consist of mainly 2 subbranch which are:

- Clustering.
- Association.

Clustering is the most applied algorithm which depends on arrange the data into groups or clusters depending on it is similarity or distances. Unsupervised ML is used generally for; Automatic labelling, similarity detection, object segmentation [38]. Some common unsupervised ML algorithms:

- K-Means Clustering.
- Apriori algorithms.
- Hierarchical clustering.
- Probabilistic clustering.
- Principal component analysis.

2.6.3 Reinforcement learning

This type of ML depends on learning from the environment and feedback provided by it. The feedback in here is a reward or punishment. The computer takes decisions and after each decision it receives a reward when the decision is true or a punishment when its decision is negative or false [38], [39]. Some common examples of this type of learning are:

- Game solving.
- Robot controlling.

3
METHODOLOGY

3.1 Data Collection

In this study known 29 different pH solutions with 0.5 interval were prepared form HCl and NaOH , the pH of each solution were examined by the use of pH strips, each solution's pH strips were captured by the use of iPhone XR camera by 2 different way; at the normal ambient lighting conditions in the laboratory also by the use of black box designed to prevent the light expect the camera flash. Overall, 6 photos were taken for each test strip (3 at ambient lighting conditions and 3 inside the black box) also the pH reference chart were captured to use in the regression as an evaluation method. After capturing images, they were transferred to PC and named orderly for each pH degree, the photos then analyzed by the use of algorithm on MATLAB to extract the most convenient regression that fit the results of analyzed images with the relevant pH value.

3.1.1 Solutions Preparing

4 main solutions with known pH degree were prepared then the rest 24 were driven from these main solutions by dilution method with 1:10 dilution factor.



Figure 3.1 Main 4 solutions prepared.

3.1.1.1 Materials and Instruments used

a. Stock solution of 37% Hydrochloric Acid

b. Volumetric flask.

c. Wash bottles.

d. Electronic scale.

e. Deionized water.

f. Beakers.

g. Conical flask.

h. Funnels.

i. Siphons.

j. Graduated cylinder.

k. Sodium hydroxide powder.

3.1.1.2 pH and solutions concentration calculating

To prepare known pH degree solutions equation 2.1 was employed to determine the desired concentration of NaOH and HCl.

i) Solutions with (0, 1, 2, 3, 4, 5 ,6) pH degree

To prepare solution with 0 pH from Stock solution of 37% Hydrochloric Acid with 12.08 M, the molarity should be calculated

$$pH = -\log [H^+] \quad (3.1)$$

$$pH = -\log [x] = 0$$

$$[x] = M\ [HCl] = 1\ M$$

So, 1 M of HCl is needed to be prepared. To prepare 1 M of HCl dilution equation was employed

$$M_1V_1 = M_2V_2 \quad (3.2)$$

Where M_1 is the concentration of 37% Hydrochloric Acid stock solution, V_1 is the volume to be used from 37% Hydrochloric Acid stock solution, M_2 is the

concentration of the desired solution which in here 1 and V_2 is the volume of water which is in here 100 ml

$$[12.08] * x = [1] * 100 \quad (3.2)$$

$$x = 8.27\ ml$$

To prepare solution with 0 pH degree the molarity should be 1 M HCl, we need 8.27 ml of 37% Hydrochloric Acid stock solution and 91.8 ml water. First, water is added to a volumetric flask, 8.3 mL of HCl is added, mixed well, and completed to 100 mL by the use of dilution method [45] using dilution factor of 1:10 we prepared solutions of (1 2 3 4 5 6) pH degree. To dilute a solution, we take 1 ml of 1 M HCl and complete it to 9 ml produce 0.1 M HCl solution with pH of 1,

$$pH = -\log[0.1] = 1 \quad (3.1)$$

The same procedure is done for pH (2 3 4 5 6), the results are summarized in the following table.

Table 3.1 (0, 1, 2, 3, 4, 5, 6) pH Solutions preparing amounts of HCl and water.

pH degree	**1**	**2**	**3**	**4**	**5**	**6**
M HCl	0.1	0.01	0.001	0.0001	0.00001	0.000001
Solution	1 ml of 1 M HCl + 9 ml water	1 ml of 0.1 M HCl + 9 ml water	1 ml of 0.01 M HCl + 9 ml water	1 ml of 0.001 M HCl + 9 ml water	1 ml of 0.0001 M HCl + 9 ml water	1 ml of 0.00001 M HCl + 9 ml water

ii) Solutions with (0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5) pH degree

To prepare solution with 0.5 pH from Stock solution of 37% Hydrochloric Acid with 12.08 M, the molarity should be calculated

$$pH = -\log [H^+] \quad (3.1)$$

$$pH = -\log [x] = 0.5$$

$$[x] = M\,[HCl] = 0.316\,M$$

So, 0.316 M of HCl is needed to be prepared. To prepare 100 ml of 0.316 M of HCl

$M_1V_1 = M_2V_2$

$$12.08 * x = 0.316 * 100 \quad (3.2)$$

$$x = 2.6\,ml$$

To prepare solution with 0.5 pH degree the molarity should be 0.316 M HCl, we need 2.6ml of 37% Hydrochloric Acid stock solution and 97.4 ml water. In the same way done at section 3.2.2.1 the solution was prepared. Also using the same dilution factor, we prepared solutions of (0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5) pH degree, to dilute the solution we take 1 ml of 0.316 M HCl and complete it to 9 ml

$$pH = -\log [0.0316] = 1.5 \quad (3.1)$$

By applying the same procedure pH of (2.5 3.5 4.5 5.5 6.5) solutions were prepared, the exact amounts are summarized in table 3.2

Table 3.2 (0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5) pH Solutions preparing amounts of HCl and water

pH degree	**1.5**	**2.5**	**3.5**	**4.5**	**5.5**	**6.5**
M HCl	0.0316	0.00316	0.000316	0.0000316	0.00000316	0.000000316
Solution	1ml of 0.316M HCl + 9 ml water	1ml of 0.0316 M HCl + 9 ml water	1ml of 0.00316 M HCl + 9 ml water	1 ml of 0.000316 M HCl + 9 ml water	1 ml of 0.0000316 M HCl + 9 ml water	1 ml of 0.00000316 M HCl + 9 ml water

iii) Solutions with (14, 13, 12, 11, 10, 9, 8) pH degree

To prepare a solution with 14 pH from sodium hydroxide salt, the molarity of NaOH solution should be calculated.

$$pH + pOH = 14 \tag{3.3}$$

$$pH - \log [NaOH] = 14$$

$$-log[NaOH] = 0$$

$$[NaOH] = 1\,M$$

To prepare 1 M NaOH solution, we have to dissolve 40.00 g of NaOH granules in 250 ml distilled water and then make up the solution to 1 Liter [45]. To prepare 100 ml of NaOH solution 4 g of sodium hydroxide salt is needed. 4 g of NaOH was weighed on a clean paper by the use of electronic scale. Then it was transferred to a beaker and dissolved in nearly 50 mL of distilled water. eventually Transferred it to the volumetric flask and completed to 100 ml. using dilution factor of 1:10 we prepared solutions of (13, 12, 11, 10, 9, 8) pH degree. To dilute a solution, we take 1 ml of 1 M NaOH and complete it to 9 ml to produce 0.1 M NaOH solution with pH of 13, the exact amounts are summarized in the following table.

Table 3.3 (14, 13, 12, 11, 10, 9, 8) pH Solutions preparing amounts of NaOH and water

pH degree	**13**	**12**	**11**	**10**	**9**	**8**
M NaOH	0.1	0.01	0.001	0.0001	0.00001	0.000001
Solution	1ml of 1M NaOH + 9 ml water	1ml of 0.1 M NaOH + 9 ml water	1ml of 0.01 M NaOH + 9 ml water	1 ml of 0.001 M NaOH + 9 ml water	1 ml of 0.0001 M NaOH + 9 ml water	1 ml of 0.00001 M NaOH + 9 ml water

iv) Solutions with (13.5, 12.5, 11.5, 10.5, 9.5, 8.5, 7.5) pH degree

To prepare solution with 13.5 pH from sodium hydroxide salt, the molarity of NaOH solution should be calculated

$$pH + pOH = 13.5 \quad (3.3)$$

$$pH - \log [NaOH] = 13.5$$

$$-log[NaOH] = 0.5$$

$$[NaOH] = 0.316\, M$$

To prepare 0.316 M NaOH solution, we have to calculate the amount of NaOH to be dissolved in 100 ml, the molar mass of NaOH [14] Na=23, O=16, H=1 overall the molar mass of NaOH is equal to 40 g/mol

$$40\ \text{g/mol} * 0.316\ \text{mol/l} = 12.64\ \text{g/l} \quad (3.2)$$

$$12.64\ \text{g/l} * 1\ \text{L}/1000\ \text{ml} * 100\ \text{ml} = 1.264\ \text{g}$$

So, we dissolved 1.264 g of NaOH granules in 100 ml distilled water to get 0.316 M NaOH solution. To dilute a solution, we take 1 ml of 0.316 M NaOH and complete it to 9 ml to produce 0.1 M NaOH solution with pH of 13, the exact amount are summarized in (Table 3.4)

Table 3.4 (13.5, 12.5, 11.5, 10.5, 9.5, 8.5, 7.5) pH Solutions preparing amounts of NaOH and water

pH degree	**13.5**	**13.5**	**13.5**	**13.5**	**13.5**	**13.5**
M NaOH	0. 316	0.0316	0.00316	0.000316	0.0000316	0.00000316
Solution	1ml of 0.316M HCl + 9 ml water	1ml of 0.0316 M HCl + 9 ml water	1ml of 0.00316 M HCl + 9 ml water	1 ml of 0.000316 M HCl + 9 ml water	1 ml of 0.0000316 M HCl + 9 ml water	1 ml of 0.00000316 M HCl + 9 ml water

v) Solution with 7 pH degrees

Distilled water was used for this pH degree.

3.2 pH Strips Immerging

A packet of 100 pH strips with 0.5-degree intervals were purchased, depending on the literature [18], the chosen pH strips have 4 different rectangular color areas that perform best accuracy.



Figure 3.2 Chosen pH strips.

One pH strip was immersed in every solution then a photo of the immersed pH strip was captured.



Figure 3.3 Samples of solutions prepared.

3.3 Photos Capturing

pH strips' photos were captured 3 times after 1,2,3 seconds of immerging in 2 different ways illustrated in the following paragraphs. For each solution 6 photos were captured overall 174 images were obtained. Also, the pH strips reference paper was captured in the 2 ways to use its data the method validation. The photos were downloaded on computer and named orderly to use it in MATLAB regression learner.

3.3.1 At laboratory ambient light

A special constant area was chosen to take pH strips photos in daytime light above a white background, the set is shown in (Figure 3.4)



Figure 3.4 Photos capturing set at laboratory ambient light.

3.3.2 Using special prepared black box

A special box made of black cork was prepared to use it as a light isolating area to capture pH strips images also above a white background with a mobile

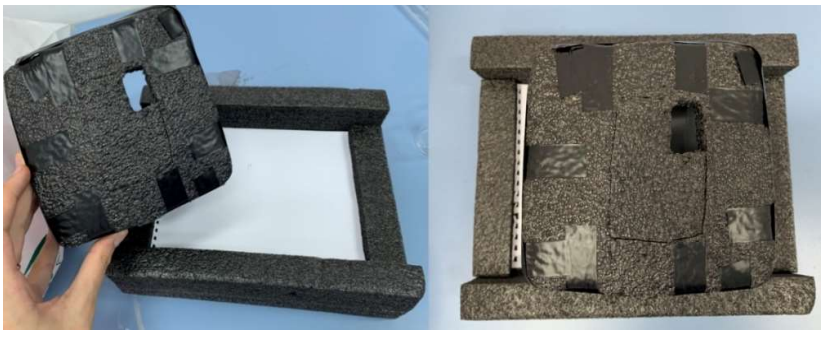


Figure 3.5 Special prepared black box.

3.4 MATLAB Coding

A code of MATLAB was written to select pixels from each color rectangle of each pH strip's photo. Pixels were manually chosen from each photo at the nearly same location. The used MATLAB code is as follows:

```
%% Clean & reload

clear, clc
%% Load images

imagefiles = dir('*.jpg');
nfiles = length(imagefiles);    % Number of files found
for ii=1:nfiles
   currentfilename = imagefiles(ii).name;
   imagesrgb{ii} = imread(currentfilename);
end
%% To Black and White and GLCM

%determine color pixel numbers
figure
for ii=1:nfiles
pixels=impixel(imagesrgb{ii})
Dataset{ii}=pixels;
end
```

Figure 3.6 MATLAB code of pixels selection.

After running the code and selecting pixels, the RGB data obtained are classified on EXCEL program for each pH degree. Then this data is called by MATLAB to process it using APP tool and Regression learner to select the most convenient model.

3.5 Model Selection

Machine learning involves a training and testing process. The system is trained first for some of the iterations then it is tested [46]. Regression learner was selected to be used since we have a supervised learning problem because our data are labelled and have input and output. 7 optimized models were chosen to be trained. The supervised trained models are as follows:

3.5.1 Linear Regression

Linear regression is the simplest form of regression, it is among the most frequently used models for machine learning [47]. It is identifying correlations between variables by suggesting that there is a linear relationship between input and output [40], linear regression is dividing to 2 subcategories.

- Simple linear regression

when having one input and one output, in other word when we are having one independent variable, simple linear regression equation is written as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.4}$$

Where y is the dependent variable, x is the dependent variable, β_0 and β_1 the equation's constant where β_0 is the interception and β_1 is the slope and ε is the random error.

- Multiple linear regression

when having more than one input variable with the desired output [48], multiple linear regression equation is written as follows:

$$\mathrm{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_{m+} \varepsilon \tag{3.5}$$

Where y is the dependent variable, x_m are the dependent variables, β_0 is the interception, β_m are the regression coefficients and ε is the error term.

By utilizing our training data set we can determine regression coefficients so we can write the equation that will be tested with the test set to be used later in prediction response for new data set.

3.5.2 Decision tree

This approach is used in both classification and regression supervised problems, it comprehends, handles categorical and quantitative data with ease, and performs well because of the tree traversal algorithm's speed. The main idea behind the decision tree is to divide the data continuously into regions according to a certain pattern, each region contains a set of data that having the same characteristic [49]. The data is divided in the nodes, each leaf node represents an outcome variable, and

the link between the nodes indicates a decision, using tree splitting logic, it moves along the tree, making choices at each node until it reaches a dead end. The decision tree's structure is shown in the following figure:

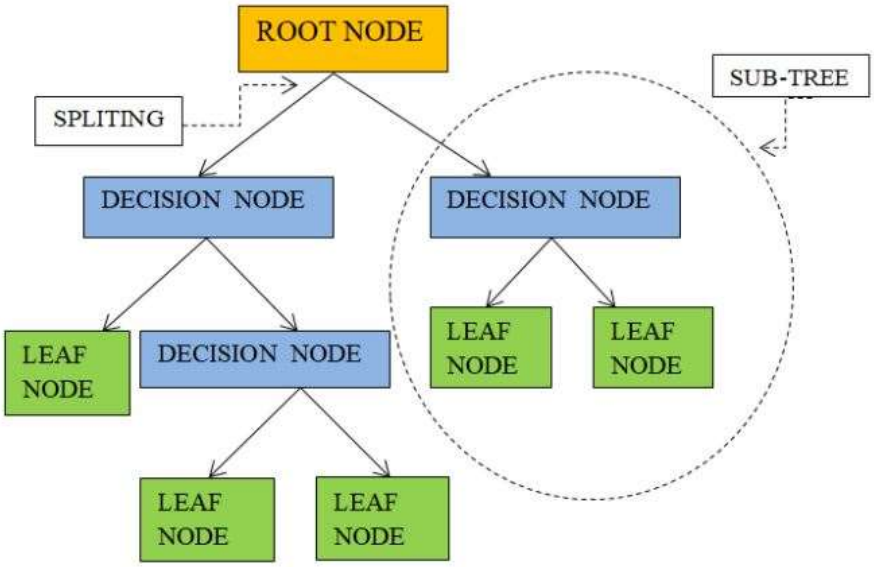


Figure 3.7 Decision tree structure [50]

The decision tree's structure is constructed using a top-down technique. Where the root node at the head symbolizes the whole training data, which is then split into two or more homogenous sets, in other word; the initial split is carried out at this node. A sub-node is referred to be a decision node when it divides into more sub-nodes. Every internal node represents a decision point that ultimately results in the outcome prediction [50].

3.5.3 Support vector machine

SVM approach is used in both classification and regression supervised problems, these machine learning models are among the most often utilized because of their high precision. This approach requires defining the hyperplane, which serves as the decision boundary. When a group of items belong to different classes, they must be separated using a decision plane [44]. In order to minimize classification error, margins are created between the classes in a way that maximizes the distance between the margin and the classes [46]. The following figure demonstrate the hyperplane, support vector and margins

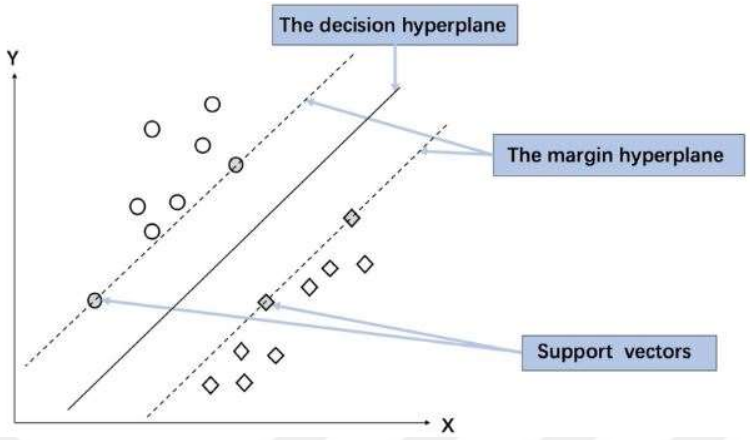


Figure 3.8 Schematic diagram of a support vector machine (SVM) [51].

when items cannot be linearly separable; in this case, the separation of objects belonging to various classes requires sophisticated mathematical operations known as kernels [47]. Which will be illustrated in the next paragraph.

3.5.4 Kernel

Kernel is a method or mathematical function that is used essentially in SVM and other algorithms. It deals with complicated mathematical functions that cannot be linearly separated, it role is to convert the incoming data into a higher-dimensional space, allowing these algorithms to carry out difficult tasks [52].

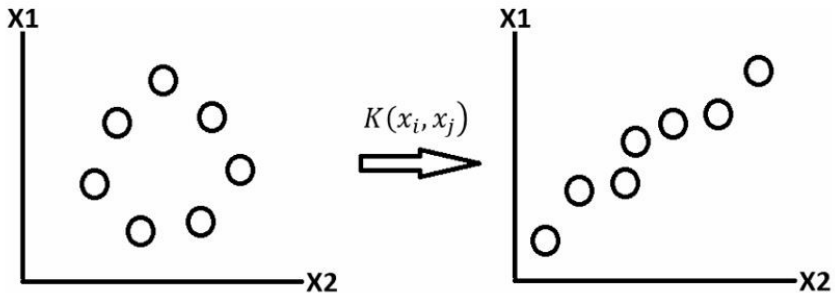


Figure 3.9 Illustration of data projection of a kernel function [53].

There are several types of kernel functions, including:

- Linear
- Polynomial
- Radial Basis function

- Sigmoid functions [54]

3.5.5 Gaussian Process Regression

GPR approach unlike its name is used for both classification and regression supervised problems, it is superior in predictions when having small data set, Gaussian processes regression models are frequently employed due to their probability, flexibility, and simplicity, making them applicable across various application areas. Since GPR models are probabilistic and provide a probability distribution of the output value instead of giving an exact forecast, it apart from other models that provide a single output [55], Gaussian process can be determined by

- The mean function m(x).
- The covariance function k(x, x').

The mean function is generally assumed to be zero. Gaussian Process can be expressed mathematically by the following equation:

$$f(x) \sim GP(m(x), k(x, x')) \tag{3.6}$$

3.5.6 Ensemble

Ensemble methods are used in both classification and regression supervised problems, these techniques are typically employed to improve the accuracy of models that are thought to have low predictive power. In order to improve prediction ability, they combine the findings of several models. There are several approaches for establishing ensembles that are applicable to a wide range of learning techniques [56], among them:

- Bayesian Voting.
- Manipulating the Training Examples.
- Manipulating the Input Features.
- Manipulating the Output Targets.
- Injecting Randomness.

Under the title of Manipulating the training examples; tow popular methods of producing ensemble are coming which are: Bagging and boosting.

3.5.6.1 Bagging

Bagging is the most fundamental method for manipulating the training set. It is called bootstrap aggregation, from which the term "bagging" originates. This type of training set is known as a bootstrap replicate of the original training set, where a training set of N training examples, randomly selected with replacement from the initial training set of m items, is sent to the learning algorithm by bagging. Each bootstrap replicate comprises, on average, 63.2% of the original training set, with multiple training examples appearing [56]. It is seen that bagging is more effective with unstable learning algorithms like decision tree and neural network when a little change in training sets lead to a large change in prediction [57].

3.5.6.2 Boosting

Boosting is a valuable technique for building multiple classifier systems. One fundamental theoretical characteristic of boosting is its capacity to lower training error, or, in other word, to boost a poor learner to become stronger, it can do that by merging many weak models to create a single strong model in order to increase prediction accuracy. Boosting trains models consecutively, concentrate on the shortcomings of the earlier models, each new model enables the ensemble to improve over time. because of this sequential training, boosting is especially successful for complicated datasets since it lowers bias and variance [56].

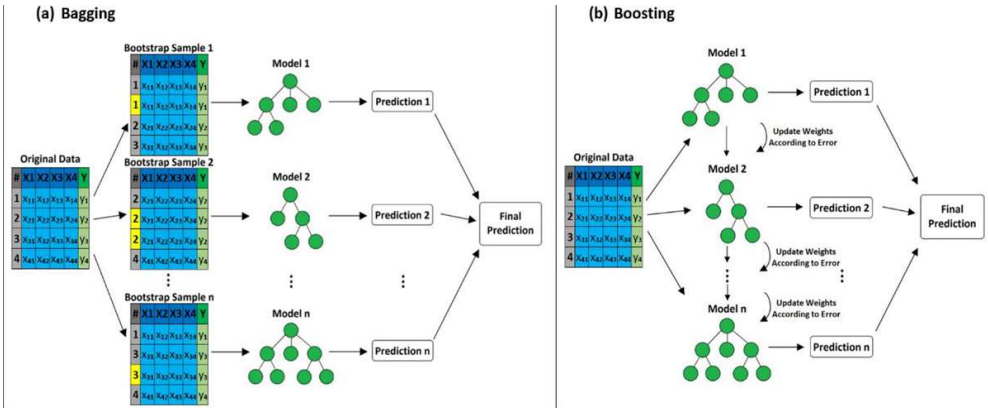


Figure 3.10 Bagging (a) and boosting (b) ensemble methods [53]

3.5.7 Neural Network

Artificial Neural Network (ANN) is a significant topic in artificial intelligence, It creates a basic model with various connections by abstracting the neural network of the human brain from information processing. ANNs are computational models that describe certain output functions by connecting a large number of nodes. Neural networks consist of multiple layers, each containing neurons, connected to form a network. A neuron processing unit in an artificial neural network can represent a variety of items, including letters, characteristics, ideas, or other significant abstraction pattern. There are three types of processing units in the network:

- Input unit: receive data from environment
- Output unit: understands the system processing result's output
- Hidden unit: hidden from view outside the system and situated between the input and output units, it could be one or more [58].

It is widely recognized for its ability to prediction non-linear variables. Similar to real neurons, artificial neural networks receive information as a stimulus, process it, and then generate a response. In a neural network, nodes in the input layer receive input data and use interconnecting connections to transmit the values to hidden layer 1. After being multiplied by weights, the results are sent to the output layer, where the goal vector is used to compute the error [59]. The following figure demonstrate Neural Network structure.

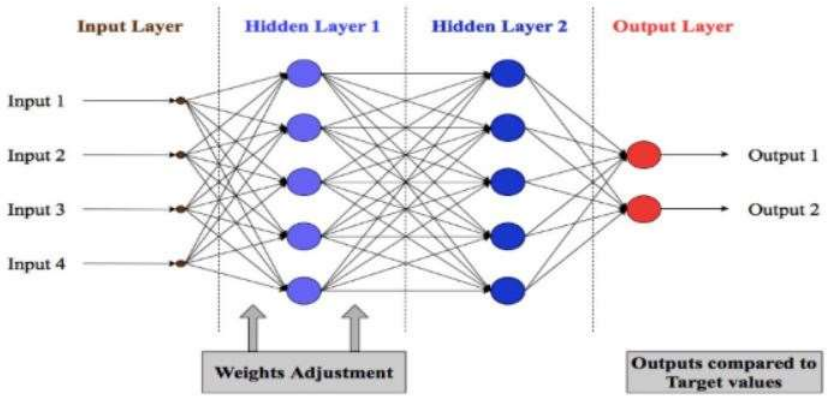


Figure 3.11 Neural Network structure [59].

3.6 Evaluation Method

Evaluation in machine learning aims to assess the effectiveness of learning algorithms on diverse data sets. After models are trained, it must be evaluated to figure out the acceptance of their performance. That is done by testing the trained models with the test set and evaluating their performance by calculating errors [60]. In order to calculate error many validation metrics are utilized such as: RMSE, MSE, R^2, MAE. In machine learning the most widely used validation methods are:

- k-fold cross validation: divides data into k subsets, tests one set, trains k models, evaluates each, and calculates the average of all errors to determine cross validation performance.
- hold-out validation: The data is divided into three sets: test, validation, and train.

The hold-out validation may cause a decrease in performance due to not considering the test set with training data so decreasing training set size. however, k-fold cross validation doesn't have this problem since it uses the whole training data. In this study k-fold cross validation will be performed.

3.6.1 Cross-validation

Cross-validation is a crucial technique in machine learning that assesses the performance and generalization capabilities of models [61]. This method depends on dividing the data set into 2 sets, one is training and the other in testing set and this is done to avoid obtaining inflated success rate [62]. This process is repeated multiple times to generate a reliable estimate of a model's performance on unseen data. Cross-validation aims to accurately simulate a model's real-world performance, identifying potential issues like overfitting or underfitting. It also minimizes biases from relying on a single train-test split and maximizes the use of available data by using each observation for both training and validation [61]. Common cross-validation methods include k-fold cross-validation, nested cross-validation, and leave-one-out cross-validation [63]. Figure 13 shows Cross validation method.

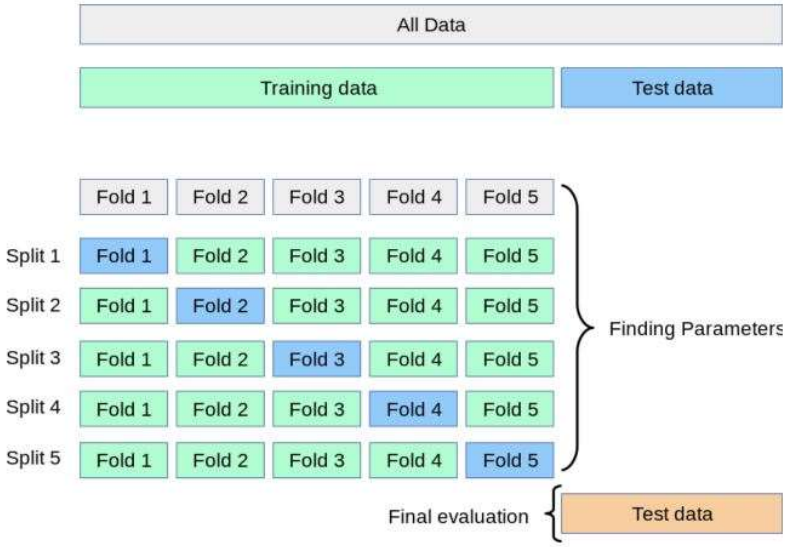


Figure 3.12 K-Fold cross validation [64].

3.6.2 Performance Metrics

As mentioned previously, some of evaluation metrics will be used in this study, in the following section theoretical information and meaning of these metrics are provided.

3.6.2.1 MSE: Mean Squared Error

Mean squared error (MSE) is one of statistical evaluation methods used in regression analysis for measuring the accuracy of a model. The MSE equation is regarded as a measure of the model's error. In other word MSE measures the differences between predicted and actual values [65]. In our study it can be characterized as the difference between the output variable that the regression model estimated and the actual output value that was acquired through experimentation. MSE can be mathematically expressed by equation (3.4)

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(P_j - A_j)^2 \quad (3.7)$$

Where, n is the number of entries within the dataset, P_j is the predicted value and A_j is the actual value [66]. To understand what the value of MSE is means it important to remember that is calculating the error of the model so minimum value of MSE represents better estimating performance.

3.6.2.2 RMSE: Root Mean Squared Error

As its name indicates RMSE is the Root of MSE the relative rankings of models are unaffected by taking the root, it produces a measure with the same units as y, it implies best fitting when it is value is close to zero since it measures the average magnitude of error in predictions [67]. MSE can be mathematically expressed by equation (3.5)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(P_j - A_j)^2} \quad (3.8)$$

Where, n is the number of entries within the dataset, P_j is the predicted value and A_j is the actual value.

3.6.2.3 R2: The Coefficient of Correlation

R^2 coefficient of correlation is a type of measurements that frequently employed for evaluating linear regression models' predictive ability, the value of R^2 in the range of 0 and 1, where 1 means perfect linear relationship which represent the perfect fit of the model, 0 means that there is no linear relationship [68]. R^2 is mathematically expressed by equation (3.6)

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2}\right] \quad (3.9)$$

Where, dominator is the total sum of squares SST, numerator is sum of squared residuals SSR, $\hat{y}_i$ is the estimated output value of the data i, and $\bar{y}_i$ is the mean value of the actual output of the data I [53].

3.6.2.4 MAE: Mean Absolute Error

Mean absolute error is a standard matrix utilized in model evaluation [67]. MAE is a statistical method for calculating the absolute value of the difference between the predicted and actual output values, in other words it is the sum of absolute error and mathematically can written as following:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(y_i - \hat{y}_{i)}\right| \quad (3.10)$$

Where y_i is the actual output value, $\hat{y}_i$ is the predicted value and n is the number of data.

3.7 Model Optimization

The last part in machine learning after constructing the models is tuning and optimizing the hyperparameters since the performance of the models is related directly to their hyperparameter settings [69], so for getting the best version of the prediction model HPO should be done. There are several methods for HPO such as: Manual testing, which is a traditional method for tuning hyper-parameters, it is require a deep comprehension of the utilized ML algorithms, however it could be difficult to apply and take a lot of time when handling complex models a huge number of data, for these reasons automatic optimization of hyper-parameters is preferable to be used [70]. In this study, Bayesian optimization will be used.

3.7.1 Bayesian Optimization

Bayesian optimization is a widely used iterative technique for HPO determination. BO uses the previously acquired findings to establish the next evaluation points. Two essential elements are used by BO to identify the subsequent hyper-parameter configuration: an acquisition function and a surrogate model. the goal of the surrogate model is fitting all of the currently observed values into the objective function. The acquisition function chooses how many points to use after acquiring the probabilistic surrogate model's prediction distribution [69].

BO procedures steps are as follows:

- Constructing probabilistic surrogate model.
- Determine the surrogate model's ideal hyper-parameter values.
- evaluate these hyper-parameter values be applying them to the real objective function.
- Add fresh findings to the surrogate model.
- repeat steps 2-4 Till the optimum number of iterations is achieved.

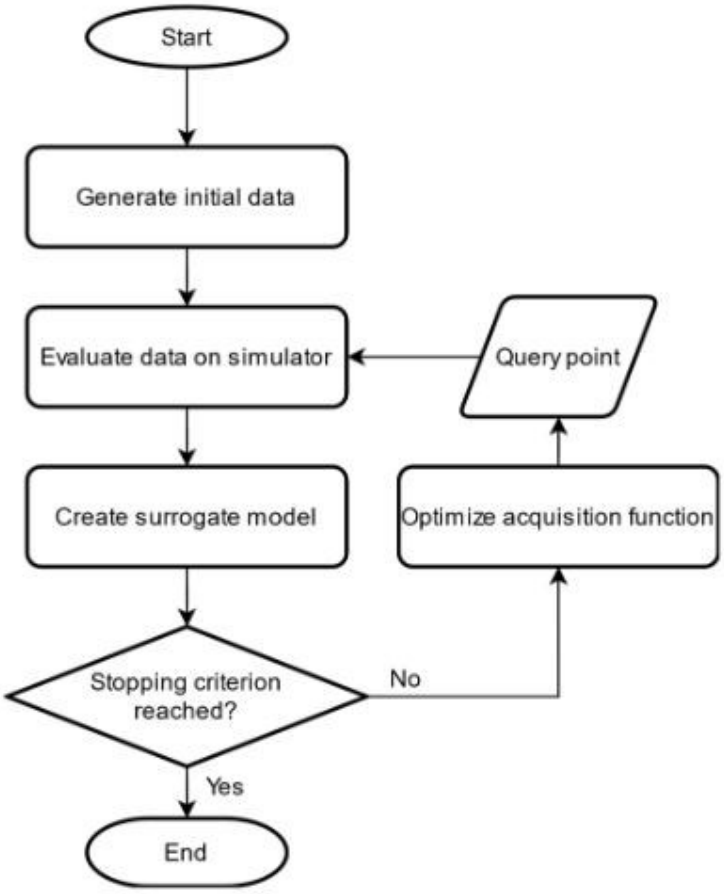


Figure 3.13 General flowchart of Bayesian optimization [70].

4
RESULTS AND DISCUSSION

In this section the pH solutions, pH strips and Basically regression models' results are discussed in detailed, since we have 2 sets of experiments the regression is done both for the 2 sets photos and as an extra evaluation method; the reference paper also was examined in the regression.

4.1 pH Solutions and pH Strips Results

29 solutions with 0.5-degree interval were prepared starting from 0 to 14 pH degree. The resulting solutions also were examined by the use of laboratory pH meter. The results given by pH meter were near to the theoretically prepared ones with some differences due to the need of calibrations of the device.

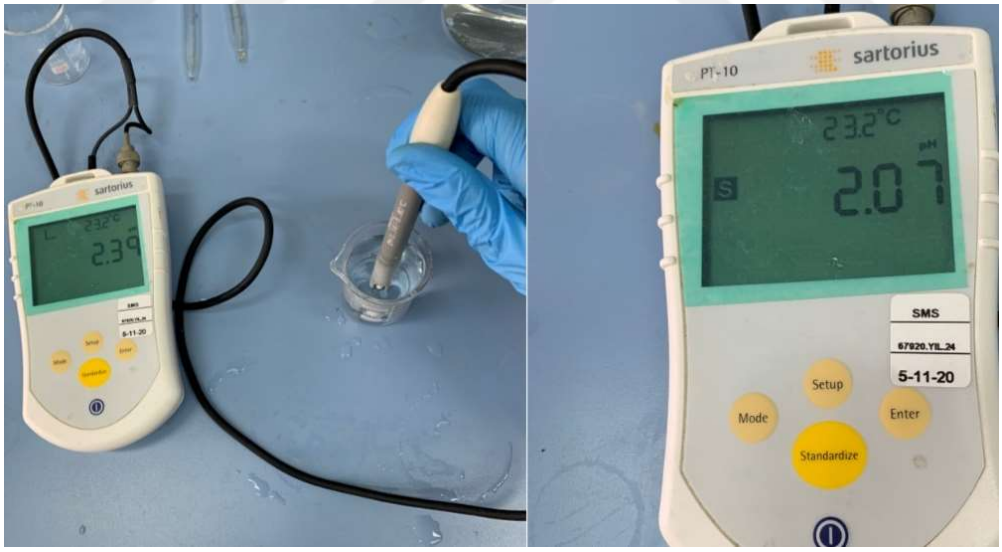


Figure 4.1 Used pH meter.

pH solutions were divided into 4 groups as illustrated in the methodology part, pH strips tested with different pH degree solutions after capturing are shown (Figure 4.2) and (Figure 4.3)

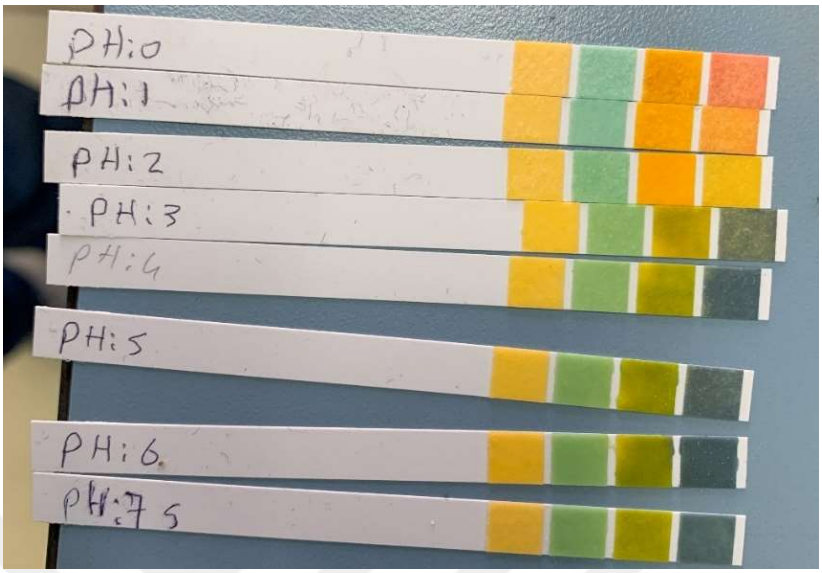


Figure 4.2 pH strips tested with (0, 1, 2, 3, 4, 5, 6, 7) pH degree solutions

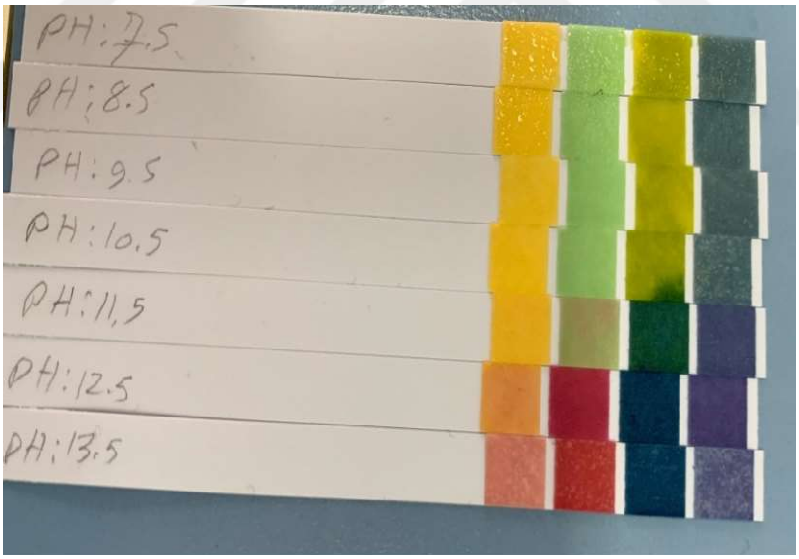


Figure 4.3 pH strips tested with (7.5, 8.5, 9.5, 10.5, 11.5, 12.5, 13.5) pH degree solutions.

It can be seen from the previous figures that there is an obvious relationship between pH strips color and pH degree changing which implies that we are on the right way and our solutions was valid. However, it is observed that the pH strip's color tends to be so close to each other around neutral pH value 7 degree, that maybe because of the dilution methods that are used in preparing the solution since the solution at

this range is very diluted and the photos are taken exactly after preparing the solution.

4.2 MATLAB GRB Results

After running the written MATLAB code, the pixels were selected manually. We obtained the RGB values for each examined pH degree, these results were transferred to EXCEL and orderly classified in tables.

Table 4.1 Example of RGB data table constructed.

	R1	G1	B1	R2	G2	B2	R3	G3	B3	R4	G4	B4	PH value
1	149	93	70	150	100	49	105	132	123	160	137	87	0
2	141	83	61	144	94	43	103	129	120	156	133	83	0
3	147	91	66	148	98	47	101	128	119	163	137	88	0
4	143	85	63	146	96	45	105	132	125	160	134	85	0
5	148	95	61	154	105	47	99	125	116	162	136	85	0.5
6	147	97	62	155	106	47	102	128	119	159	136	86	0.5
7	150	98	61	155	107	45	99	125	116	161	135	84	0.5
8	150	97	63	152	101	44	102	128	119	182	156	107	0.5
9	159	113	64	165	116	49	102	132	120	168	142	85	1
10	158	112	63	162	113	46	101	131	119	174	148	91	1
11	159	113	64	166	115	49	98	128	118	163	136	81	1
12	157	111	62	165	114	48	99	129	117	168	141	86	1
13	151	114	61	164	115	46	92	123	108	169	144	90	1.5
14	154	117	64	165	118	48	92	123	107	165	140	86	1.5
15	151	114	61	167	118	49	93	124	109	165	140	86	1.5
16	153	116	63	163	114	45	93	124	109	162	137	83	1.5
17	185	151	80	196	141	50	119	154	134	183	157	100	2
18	184	152	79	196	141	51	116	151	131	179	154	98	2
19	167	135	74	174	123	44	98	127	109	176	150	93	2
20	165	133	72	176	127	48	108	137	119	179	154	98	2
21	106	95	67	150	115	47	99	123	99	152	129	75	2.5
22	104	93	65	145	112	43	100	124	98	163	138	84	2.5
23	114	103	75	152	118	47	97	121	95	165	140	86	2.5
24	111	100	70	145	112	43	97	121	95	159	136	82	2.5
25	95	88	80	113	97	38	99	123	97	160	138	81	3
26	91	87	78	115	98	42	101	125	99	163	140	86	3
27	102	95	87	113	98	41	98	122	96	164	139	82	3

The data then is imported from EXCEL and used in the suggested regression modules.

4.3 Regression Learner

The data was imported, and since the data are labelled including output, input and our variable is continuous; supervised regression machine learning modules were chosen to be trained from More APPS tool on MATLAB, machine learning toolbox, Regression learner.

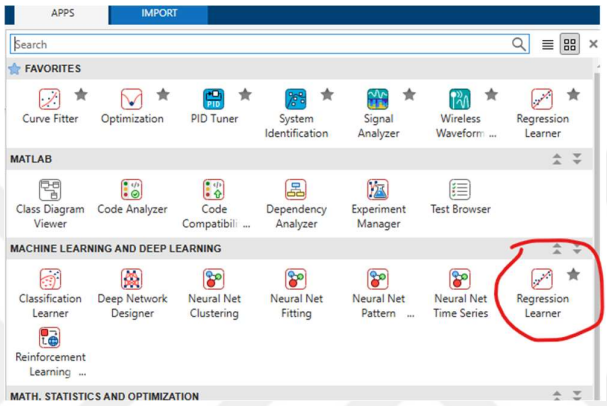


Figure 4.4 MATLAB regression learner.

cross validation method with 10-fold, and Hyperparameter; Bayesian optimization with 30 iterations were used in models' regression.

4.3.1 Chosen machine learning models

7 modules have been selected to be examined for our dataset.

Table 4.2 Machine learning examined models.

#	**Model**
1	Tree; optimizable tree
2	SVM; optimizable SVM
3	Multiple; all linear
4	Gaussian Process Regression; optimizable
5	Kernel
6	Ensemble
7	Neural Network

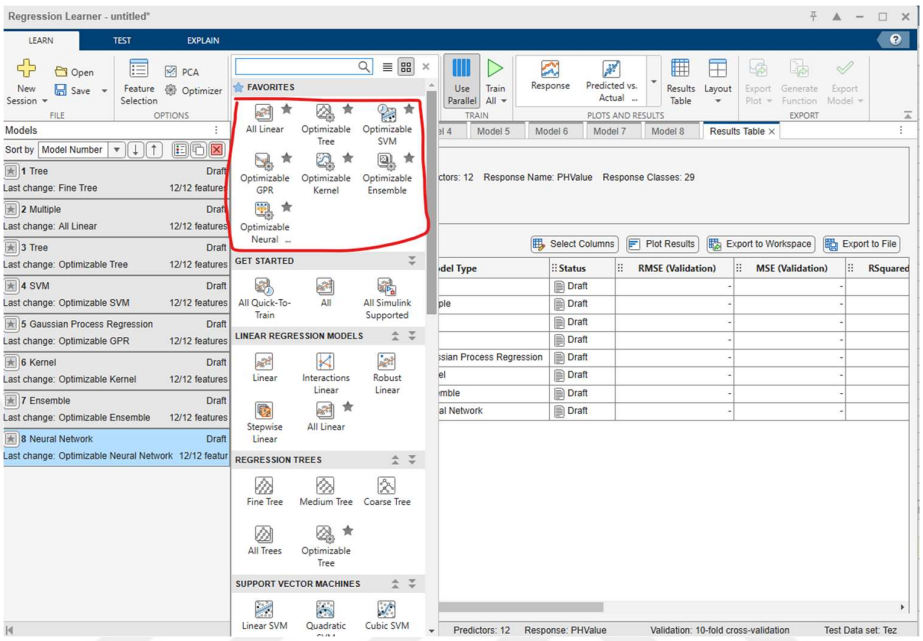


Figure 4.5 Selected regression models used.

4.4 Regression Results

Regression results will be discussed in the following sections starting with the reference paper regression results. Reference paper of pH strips was captured and its RGB values were processed with the selected algorithms to examine the more efficient set of experiments to be used. Also, it is regarded as a validation method that assures the information given by literature about the presence of a linear relationship between pH degree and color changing. At this part and before processing our experimental data, observation of effects of number of pixels on the accuracy of the model were conducted. Detecting that the more the number of pixels we are feeding the algorithms the more accurate model we get.

4.4.1 Reference paper inside the black box

Starting with 2 pixels selected from each color rectangular of each photo, 8 pixels from each photo, overall, 232 pixels RGB values were fed to the regression learner algorithms. The results of validation evaluation metrics are given in the table below:

Table 4.3 Validation evaluation of reference paper inside the black box with 232 pixels RGB values feed.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	0.223182	0.04981	0.997245	0.173941
2	Interactions Linear	1.333328	1.777764	0.901685	0.922035
3	Robust Linear	0.248181	0.061594	0.996594	0.188703
4	Stepwise Linear	0.232814	0.054202	0.997002	0.177698
5	Tree; optimizable tree	0.994219	0.988472	0.945335	0.785249
6	SVM; optimizable SVM	0.693285	0.480644	0.973419	0.492324
7	GPR; optimizable	0.220419	0.048585	**0.997313**	0.176836
8	Kernel	2.090851	4.371659	0.758236	1.578782
9	Ensemble	0.722178	0.521541	0.971157	0.501132
10	Neural Network	0.222842	0.049659	0.997254	0.173856

Regression results were the same as the information given by literature, there is a strong linear relationship between our variables. It is seen from the table above; GPR regression model is the most accurate model with 0.997313 R^2 also NN and linear regression gives good prediction results. For better understanding of the relationship the following plot is given:

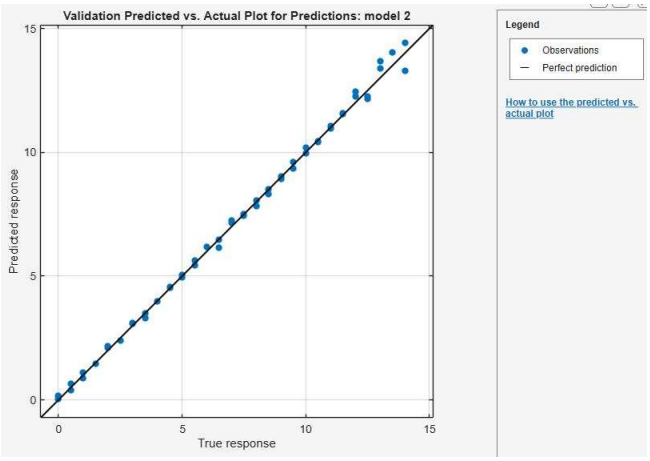


Figure 4.6 GPR plot for validation predicted vs. actual with 232 pixels RGB values feed.

It can be understood from figure 16 that the predicted data are near to the actual ones. The number of pixels was duplicated by selecting 16 pixels from each photo and overall, 464-pixel RGB values in order to raise the accuracy of the models, the results are as follows:

Table 4.4 Validation evaluation of reference paper inside the black box with 464 pixels RGB values feed.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	0.190722	0.036375	0.997932	0.155031
2	Interactions Linear	0.25945	0.067314	0.996173	0.178232
3	Robust Linear	0.1934	0.037404	0.997873	0.156339
4	Stepwise Linear	0.137548	0.01892	**0.998924**	0.105737
5	Tree; optimizable tree	0.503856	0.253871	0.985567	0.398095
6	SVM; optimizable SVM	0.204422	0.041788	0.997624	0.162927
7	GPR; optimizable	0.147494	0.021755	0.998763	0.112329
8	Kernel	1.237472	1.531337	0.912938	0.9475
9	Ensemble	0.276898	0.076673	0.995641	0.179984
10	Neural Network	0.190524	0.036299	0.997936	0.154911

A slightly incraesing in accuracy were obsereved as the number o pixels is increased, again the number of pixels has raised to be 24 pixel from each photo, overall, 696 pixels RGB values, the results are recorded in the following table:

Table 4.5 Validation evaluation of reference paper inside the black box with 696 pixels RGB values feed.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	0.187042	0.034985	0.998021	0.148869
2	Interactions Linear	0.185975	0.034587	0.998043	0.112591
3	Robust Linear	0.188337	0.035471	0.997993	0.149109
4	Stepwise Linear	0.132718	0.017614	0.999004	0.098747

Table 4.6 Validation evaluation of reference paper inside the black box with 696 pixels RGB values feed (continued).

5	Tree; optimizable tree	0.458222	0.209967	0.988121	0.275969
6	SVM; optimizable SVM	0.187527	0.035166	0.998011	0.148214
7	GPR; optimizable	0.106884	0.011424	**0.999354**	0.081172
8	Kernel	1.237579	1.531602	0.913352	0.926889
9	Ensemble	0.290437	0.084354	0.995228	0.189843
10	Neural Network	0.188644	0.035587	0.997987	0.153315

Again GPR is given the most accurate results besides with other models that are also gives a high accuracy. Figure (4.7) shows GPR validation plot for predicted vs actual predictions, it can be compared to Figure (4.6) and find out that the GPR model gives perfect fitting.

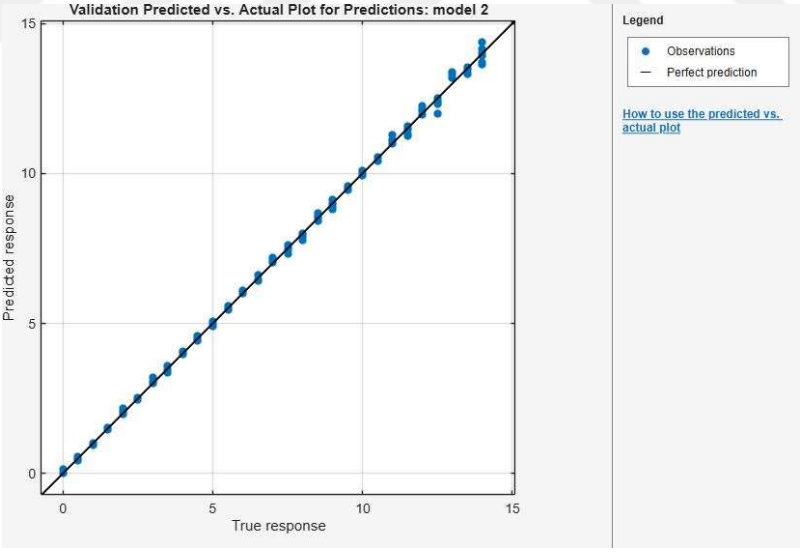


Figure 4.7 GPR plot for validation predicted vs. actual with 696 pixels RGB values feed.

a higer accuracy is observed as the number of pixels is increased, so it is decieded to select 24 from each photo, it could be more than this number ofcourse but since the pixels are selecting manulay it is become hard to select a huge number of pixel,

maybe in future reserches we can make a program for selecting the pixels automatically and having higher number of data and resulting a higher accuracy.

4.4.2 Reference paper in ambient laboratory light

In this section as mentioned before it is decided to select 24 pixels from each photo. So, here we are examining the effect of the set of 2 experiments in having more accurate results. The following table shows the evaluation metrics of reference paper inside the black box with 696 pixels RGB values feed.

Table 4.7 Validation evaluation of reference paper in ambinet laboratory light with 696 pixels RGB values feed.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	0.368296	0.135642	0.992325	0.238059
2	Interactions Linear	0.447067	0.199869	0.98869	0.252655
3	Robust Linear	0.477305	0.22782	0.987109	0.258227
4	Stepwise Linear	0.360548	0.129995	**0.992644**	0.20331
5	Tree; optimizable tree	0.533025	0.284116	0.983923	0.301456
6	SVM; optimizable SVM	0.408498	0.16687	0.990557	0.251096
7	GPR; optimizable	0.368296	0.135642	**0.992325**	0.238059
8	Kernel	0.447067	0.199869	0.98869	0.252655
9	Ensemble	0.477305	0.22782	0.987109	0.258227
10	Neural Network	0.360548	0.129995	**0.992644**	0.20331

Nearly the same results are obtained with using the ambient laboratory light set of experiment with very slight difference in accuracy between GPR and NN algorithms. It is seen that the results of experiments set inside the box are a bit more accurate than the second set. However, the 2 sets ‘results will be studied.

4.4.3 Experiment Set Inside the Black Box Results

For this section and the following one we have 2 photos of each pH strip to extract data from so the number of pixels used is duplicated to be 1392 pixels. The regression results are shown in the following table:

Table 4.8 Validation evaluation metrics of reference paper inside the black box with 1392 pixels RGB values feed.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	1.633231	2.667443	0.848303	1.289426
2	Interactions Linear	2.060933	4.247444	0.758448	1.282451
3	Robust Linear	1.642961	2.69932	0.84649	1.290046
4	Stepwise Linear	1.588182	2.522322	0.856556	1.166559
5	Tree; optimizable tree	1.532539	2.348677	0.866431	0.955648
6	SVM; optimizable SVM	1.627308	2.64813	0.849401	1.269415
7	GPR; optimizable	1.269285	1.611083	**0.908378**	0.801138
8	Kernel	2.059148	4.240092	0.758866	1.541281
9	Ensemble	1.368061	1.871591	0.893563	0.90806
10	Neural Network	1.473782	2.172035	0.876477	1.027211

From the table above, as can be predicted the best model to fit our data is GPR with R^2 of 0.9083278, the accuracy here is not above 0.9 because of having our intermediate solution pH degree to be close to each other as we have seen in section (4.1). pH strips result photos, near the 7-degree pH the solutions becoming as the same color. It is also can be understood from the plot of responses for predictions.

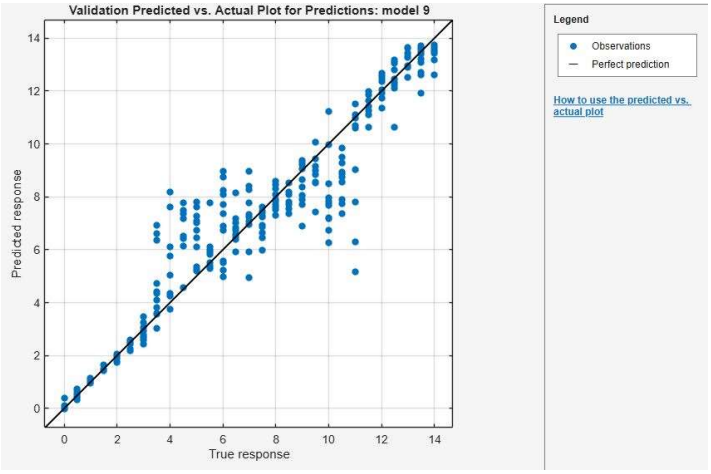


Figure 4.8 GPR plot for validation predicted vs. actual with 1392 pixels RGB values feed.

The reason for having intermediate divergence and outliers in results me be because using dilution method that the solutions near 7 pH degree are over diluted and become close to each other, but since the results are changing linearly it is not regarded as a problem because we still can construct model with good efficiency.

4.4.4 Experiment Set in Ambient Laboratory Light Results

This is the last experiment, done under ambient laboratory light. As done in part (4.4.3). 1392 pixels RGB values are fed to the algorithms models and the results of training are listed below (Table 4.8).

Table 4.9 Validation evaluation metrics of reference paper in ambient laboratory light with 1392 pixels RGB values feed.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	1.298562	1.686263	0.903328	1.03658
2	Interactions Linear	1.17662	1.384436	0.920631	0.819431
3	Robust Linear	1.302539	1.696608	0.902735	1.036456
4	Stepwise Linear	1.474904	2.175342	0.875289	0.910907
5	Tree; optimizable tree	0.954069	0.910247	0.947816	0.466812
6	SVM; optimizable SVM	1.291364	1.667621	0.904397	0.92207
7	GPR; optimizable	0.668345	0.446685	**0.974392**	0.390423
8	Kernel	1.79955	3.238382	0.814346	1.294246
9	Ensemble	0.872016	0.760412	0.956406	0.464291
10	Neural Network	0.677955	0.459623	**0.97365**	0.394488

The results show good prediction. The most accurate model is the same again, GPR with R^2 of 0.9743, Also NN gives very good prediction results with R^2 0.97365, the following figures show the validation predicted vs. actual predictions of GPR and NN.

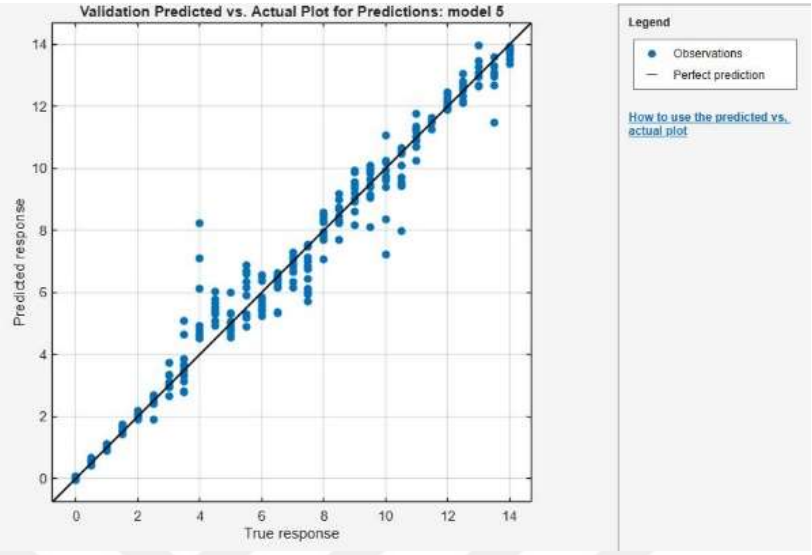


Figure 4.9 GPR plot for validation predicted vs. actual with 1392 pixels RGB values feed.

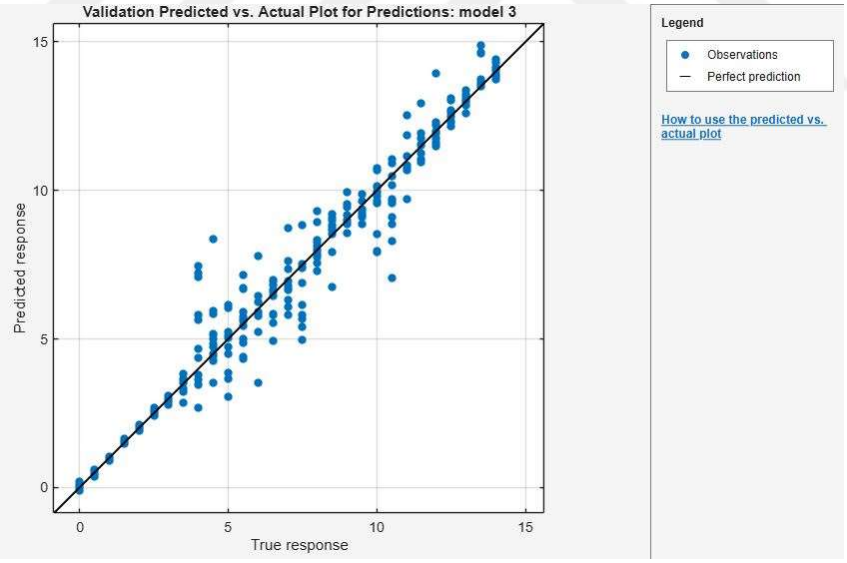


Figure 4.10 NN plot for validation predicted vs. actual with 1392 pixels RGB values feed.

Since 10-fold cross validation method is utilized in this study, a set of 10% presents of whole data was examined after training, test set results are given (Table 4.9) showing that the constructed model fits the data.

Table 4.10 Test set results.

#	Model type	RMSE	MSE	R^2	MAE
1	Linear	1.18306	1.399631	0.917908	1.013346
2	Interactions Linear	1.175546	1.381908	0.918947	0.960924
3	Robust Linear	1.191091	1.418698	0.91679	1.01972
4	Stepwise Linear	1.09599	1.201195	0.929547	0.854367
5	Tree; optimizable tree	1.181228	1.3953	0.918162	0.622041
6	SVM; optimizable SVM	1.150085	1.322696	0.92242	0.910473
7	GPR; optimizable	0.760998	0.579118	**0.966033**	0.467275
8	Kernel	1.2029	1.446969	0.915131	0.984235
9	Ensemble	1.03484	1.070894	0.937189	0.648234
10	Neural Network	1.1099	1.2318	0.93141	0.71093

Test predicted vs actual predictions are shown (Figure 4.11)

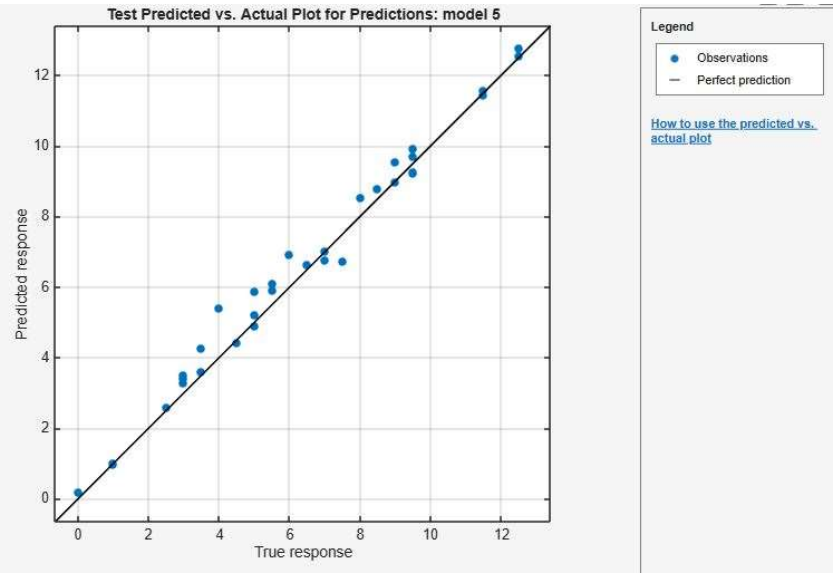


Figure 4.11 Test predicted vs actual predictions.

4.5 Regressions Findings

After trying the 2 sets of experiments and depending on the results obtained, it is obvious that capturing pH strips images under the ambient laboratory light given more precise regression results that maybe because of the contrast of the surrounding white environment and pH strips rectangular colors which make colors more distinguishable while capturing the pH strips photos inside the black box with just camera flash light results dark environment which can affect the RGB values. So, it is chosen to construct regression prediction model depending on section (4.4.4) results by chosen Gaussian process regression (GPR) model. In the following table model optimization results are given

Table 4.11 Optimzation results.

Hyperparameter	**Value**
Sigma	3.8404
Basis function	Zero
Kernel function	Isotropic Exponential
Kernel scale	491.3536
Standardize data	false

4.6 Validation with the test experiment set

As mentioned previously one set of data was kept in order to examine the model, the results are shown (Figure 4.11), the results impalys that the model gives acceptable prediction results.

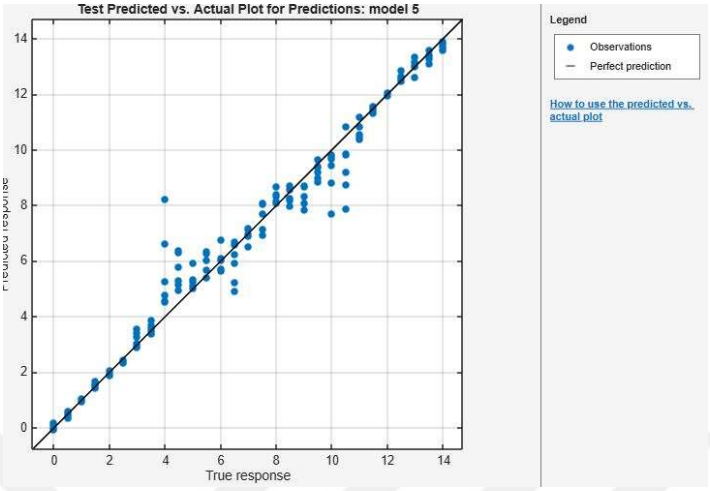


Figure 4.12 Plot of validation results for predicted vs. actual.

Test set performance metrics are given (Table 4.11)

Table 4.12 Test set performance metrics.

Performance Metric	Value
RMSE	0.657
MSE	0.43165
R^2	0.97533
MAE	0.36463

4.7 Comparison with other study

In this part, a comparison with other relevant study from literature that uses a very close approach is conducted. The results are summarized in Table (4.9) showing that Gaussian process regression (GPR) algorithm is doing best with our data while K-Nearest Neighbor (KNN) is doing best with study's data knowing that GRP algorithm is used for regression analysis and KNN is used for classification.

Table 4.13 Compression between our study and relevant study from literature.

		[71]				**Our study**			
Data size		2689 RGB values				1392 RGB values			
Validation method		5-fold cross-validation				10-fold cross-validation			
Validation metrics		**RMSE**	**MSE**	**R^2**	**MAE**	**RMSE**	**MSE**	**R^2**	**MAE**
1	**GPR**	☒	☒	☒	☒	0.6683	0.4466	0.9743	0.3904
2	**SVM**	1.893	3.584	0.744	1.460	1.2913	1.6676	0.9043	0.9220
3	**NN**	0.891	0.794	0.943	0.578	0.6779	0.4596	0.9736	0.3944
4	**LR**	2.097	4.399	0.686	1.637	1.2985	1.6862	0.9033	1.0365
5	**KNN**	0.266	0.071	0.995	0.160	☒	☒	☒	☒

4.8 Future Aspects

This study can be expanded in the future with more sophisticated experiments circumferences, a greater number of input data which can be selected automatically by preprepared program to construct a mobile phone (iPhone) application that automatically determines pH value of the strips just by captured image.

5
CONCLUSION

In this study, pH value determination was performed depending on analyzing pH strips captured images by the use of mobile phone camera using different machines learning algorithms on MATLAB. The determination was obtained for 29 different pH value solutions prepared using NaOH and HCl with known pH value with 0.5-degree difference. 2 sets of experiments were used to investigate the pH value by the use of 4 rectangular pH strips: one by the ambient laboratory light and the other by the use of special prepared black box to isolate the light and access just the flash camera light. pH strips images were captured by the previously mentioned 2 ways immediately after dipping in the solutions and 3 images were captured for each strip overall 174 images. Also, the reference paper was captured in the 2 mentioned ways to use it as a validation methos. MATLAB was used to select pixels from each image at relativly same location. pH reference paper was used to decide the number of pixels and the better experiment set, Number of pixels decided to be 24 from each photo overall 1392 pixels, one data set were kept for testing the model. The colorimetric analysis of the selected pixels done and the RGB values were extracted by the use of MATLAB code. The RGB values were transferred to EXCEL Microsoft to create a dataset for pH value versus RGB values. This dataset is then called by MATLAB to use it in constructing a model that eventually determines the pH value from captured images of pH strip. A 10-fold cross-validation method was used to validate the performance of the models. It was found that the most appropriate model is Gaussian Process Regession (GPR) with 0.974392 R^2. It can be concluded from this study that machine learning algorithms can determine the pH value of strips more accurately than naked person eyes. So, the model suggested by this study is recorded as a successful model for pH determination. This study can be expanded in the future to construct a mobile phone (iPhone) application that automatically determines pH value of the strips just by captured image. The study's findings suggests that machine learning algorithms are effective tools for producing reliable models for estimating pH value.

REFERENCES

[1] H. Li, X. Wang, X. Li, and H.-Z. Yu, "Quantitative pH Determination Based on the Dominant Wavelength Analysis of Commercial Test Strips," *Anal Chem*, vol. 93, no. 46, pp. 15452–15458, Nov. 2021, doi: 10.1021/acs.analchem.1c03393.

[2] F. Syaikhoni Aziz, A. Prasetya Wibawa, I. Made Wirawan, A. Iskandar Syah, and T. Huy Duy, "Image Processing System for pH Classification Using Biosensors," 2021.

[3] D. Zhao, Z. Hao, J. Wang, and J. Tao, "Effects of pH in irrigation water on plant growth and flower quality in herbaceous peony (Paeonia lactiflora Pall.)," *Sci Hortic*, vol. 154, pp. 45–53, May 2013, doi: 10.1016/j.scienta.2013.02.023.

[4] John Brennan, "Potential Sources of Error Using pH Strips," https://sciencing.com/potential-error-using-ph-strips-8582444.html.

[5] M. G. Hur *et al.*, "Development of a vision-based pH reading system," *Journal of the Korean Physical Society*, vol. 67, no. 8, pp. 1448–1453, Oct. 2015, doi: 10.3938/jkps.67.1448.

[6] N. K. Vuong, S. Chan, and C. T. Lau, "Classification of pH levels using a mobile phone," in *2009 IEEE 13th International Symposium on Consumer Electronics*, IEEE, 2009, pp. 823–827. doi: 10.1109/ISCE.2009.5156811.

[7] A. Beycıoğlu, B. Çomak, and D. Akçaabat, "Evaluation of pH Value by Using Image Processing," *Acta Phys Pol A*, vol. 132, no. 3–II, pp. 1142–1144, Sep. 2017, doi: 10.12693/APhysPolA.132.1142.

[8] S. D. Kim, Y. Koo, and Y. Yun, "A Smartphone-Based Automatic Measurement Method for Colorimetric pH Detection Using a Color Adaptation Algorithm," *Sensors*, vol. 17, no. 7, p. 1604, Jul. 2017, doi: 10.3390/s17071604.

[9] J. Li *et al.*, "Mobile App to Quantify pH Strips and Monitor Titrations: Smartphone-Aided Chemical Education and Classroom Demonstrations," *J Chem Educ*, vol. 100, no. 9, pp. 3634–3640, Sep. 2023, doi: 10.1021/acs.jchemed.3c00227.

[10] R. J. Myers, "One-hundred years of pH," *J Chem Educ*, vol. 87, no. 1, pp. 30–32, Jan. 2010, doi: 10.1021/ed800002c.

[11] J. Mitschele, "Chemistry: The Central Science, Sixth Edition (Brown, Theodore L.; LeMay, H. Eugene, Jr.; Bursten, Bruce E.)," *J Chem Educ*, vol. 72, no. 5, 1995, doi: 10.1021/ed072pa107.2.

[12] J. Schleicher, “Information on pH measurement.” Accessed: Jun. 23, 2024. [Online]. Available: https://www.manoraz.com/_Uploads/dbsAttachedFiles/FAS622gb_pH.pdf

[13] K. Addy, L. Green, and E. Herron, “Water quality standards for pH?,” 2004. [Online]. Available: www.uri.edu/ce/wq/has/html/has_wellfacts.html

[14] D. W. Ball, J. A. Key, and B. Victoria, “Introductory Chemistry - 1st Canadian Edition,” 2011.

[15] J. G. Webster, *The measurement, instrumentation, and sensors handbook*. CRC Press, 1999.

[16] W. Vonau and U. Guth, “pH Monitoring: a review,” *Journal of Solid State Electrochemistry*, vol. 10, no. 9, pp. 746–752, Sep. 2006, doi: 10.1007/s10008-006-0120-4.

[17] J. Franzke, C. Meyer, S. Müller, T. Krähling, and A. Michels, “Microdischarges for Analytical Atomic Spectrometry: Design Considerations and Applications,” in *Encyclopedia of Analytical Chemistry*, Wiley, 2011. doi: 10.1002/9780470027318.a9187.

[18] N. A. Metheny, E. M. Gunn, C. S. Rubbelke, T. F. Quillen, U. R. Ezekiel, and K. L. Meert, “Effect of pH Test-Strip Characteristics on Accuracy of Readings,” *Crit Care Nurse*, vol. 37, no. 3, pp. 50–58, Jun. 2017, doi: 10.4037/ccn2017199.

[19] A. Watertest Stsystems is based in South West Sydeny, “History of pH strips.”

[20] K. K. Daruwala C Patel and K. Viswanathan Baskaran, “Development of Colour Chart Analysis in Mobile Camera System for pH Detection using Litmus Paper.” [Online]. Available: https://www.researchgate.net/publication/379337923

[21] A. Aril, M. Masriani, and H. A. Melati, “Development of pH Measuring Strip Test Based On West Kalimantan Local Plant,” *EduChemia (Jurnal Kimia dan Pendidikan)*, vol. 7, no. 2, p. 267, Jul. 2022, doi: 10.30870/educhemia.v7i2.15252.

[22] M. Leung, L. Zhang, X. Li, and H.-Z. Yu, “Superhydrophobic Paper Strips with Embedded Agarose-Anthocyanin Mini-Discs for Point-of-Need Quantitative pH Measurements,” *Anal Chem*, vol. 96, no. 39, pp. 15808–15815, Oct. 2024, doi: 10.1021/acs.analchem.4c04242.

[23] S. Hoving, B. Gerrits, H. Voshol, D. Müller, R. C. Roberts, and J. van Oostrum, “Preparative two-dimensional gel electrophoresis at alkaline pH using narrow range immobilized pH gradients,” *Proteomics*, vol. 2, no. 2, pp. 127–134, Feb. 2002, doi: 10.1002/1615-9861(200202)2:2<127::AID-PROT127>3.0.CO;2-Y.

[24] “mage Processing Algorithms.”

[25] A. Rosenfeld, “Image Processing and Recognition,” 1979, pp. 1–57. doi: 10.1016/S0065-2458(08)60581-1.

[26] Rohit Kundu, “Image Processing: Techniques, Types, & Applications.”

[27] H. Basak, R. Kundu, and R. Sarkar, “MFSNet: A multi focus segmentation network for skin lesion segmentation,” *Pattern Recognit*, vol. 128, p. 108673, Aug. 2022, doi: 10.1016/j.patcog.2022.108673.

[28] M. Oberholzer, M. □streicher, H. Christen, and M. Br□hlmann, “Methods in quantitative image analysis,” *Histochem Cell Biol*, vol. 105, no. 5, pp. 333–355, May 1996, doi: 10.1007/BF01463655.

[29] M. Alauddin, S. Ahmed, R. Akter, Md. A. Ullah, and Md. A. Mannan, “A Study on the Development of Touchless Mouse by Object Tracking through Webcam,” *International Journal of Internet and Distributed Systems*, vol. 05, no. 01, pp. 1–17, 2023, doi: 10.4236/ijids.2023.51001.

[30] By Simplilearn, “What Is Image Processing : Overview, Applications, Benefits, and More,” Aug 13, 2024.

[31] Rohit Kundu, “Image Processing: Techniques, Types, & Applications [2024].”

[32] B. Mahesh, “Machine Learning Algorithms - A Review,” *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/ART20203995.

[33] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, “What is Machine Learning? A Primer for the Epidemiologist,” *Am J Epidemiol*, Oct. 2019, doi: 10.1093/aje/kwz189.

[34] A. I. Khan and S. Al-Habsi, “Machine Learning in Computer Vision,” *Procedia Comput Sci*, vol. 167, pp. 1444–1451, 2020, doi: 10.1016/j.procs.2020.03.355.

[35] J. Wei *et al.*, “Machine learning in materials science,” *InfoMat*, vol. 1, no. 3, pp. 338–358, Sep. 2019, doi: 10.1002/inf2.12028.

[36] P. Langley and H. A. Simon, “Applications of machine learning and rule induction,” *Commun ACM*, vol. 38, no. 11, pp. 54–64, Nov. 1995, doi: 10.1145/219717.219768.

[37] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput Sci*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[38] Giuseppe Bonaccorso, *Machine Learning Algorithms Popular algorithms for data science and machine learning*, 2nd ed. MUMBAI: PACKT, 2018.

[39] T. Jo, *Machine Learning Foundations*. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-65900-4.

[40] V. Nasteski, “An overview of the supervised machine learning methods,” *HORIZONS.B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/HORIZONS.B.04.1.17.P05.

[41] John Terra, “Regression vs. Classification in Machine Learning for Beginners.”

[42] Aized Amin Soofi and Arshad Awan, “Classification Techniques in Machine Learning: Applications and Issues,” *Journal of Basic & Applied Sciences*, vol. 13, pp. 459–465, Jan. 2017, doi: 10.6000/1927-5129.2017.13.76.

[43] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, “Logistic Regression for Machine Learning in Process Tomography,” *Sensors*, vol. 19, no. 15, p. 3400, Aug. 2019, doi: 10.3390/s19153400.

[44] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” *J Phys Conf Ser*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

[45] “Molarity ‘Your Safer Source for Science Supplies’ Laboratory Solution Preparation,” 2011.

[46] Y. Kumar, K. Kaur, and G. Singh, “Machine Learning Aspects and its Applications Towards Different Research Areas,” in *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, IEEE, Jan. 2020, pp. 150–156. doi: 10.1109/ICCAKM46823.2020.9051502.

[47] S. Ray, “A Quick Review of Machine Learning Algorithms,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.

[48] S. Rong and Z. Bao-wen, “The research of regression model in machine learning field,” *MATEC Web of Conferences*, vol. 176, p. 01033, Jul. 2018, doi: 10.1051/matecconf/201817601033.

[49] F. ELMAZ, Ö. YÜCEL, and A. Y. MUTLU, “Machine learning based approach for predicting of higher heating values of solid fuels using proximity and ultimate analysis,” *International Journal of Advances in Engineering and Pure Sciences*, vol. 32, no. 2, pp. 145–151, Jun. 2020, doi: 10.7240/jeps.558378.

[50] S. Bishnoi and B. K. Hooda, “Decision Tree Algorithms and their Applicability in Agriculture for Classification,” *Journal of Experimental Agriculture International*, pp. 20–27, May 2022, doi: 10.9734/jeai/2022/v44i730833.

[51] Z.-M. Zou, D.-H. Chang, H. Liu, and Y.-D. Xiao, “Current updates in machine learning in the prediction of therapeutic outcome of hepatocellular carcinoma: what should we know?,” *Insights Imaging*, vol. 12, no. 1, p. 31, Dec. 2021, doi: 10.1186/s13244-021-00977-9.

[52] M. C. Jones, S. J. Davies, and B. U. Park, “Versions of Kernel-Type Regression Estimators,” *J Am Stat Assoc*, vol. 89, no. 427, pp. 825–832, Sep. 1994, doi: 10.1080/01621459.1994.10476816.

[53] H. Yaka, M. A. Insel, O. Yucel, and H. Sadikoglu, “A comparison of machine learning algorithms for estimation of higher heating values of biomass and fossil fuels from ultimate analysis,” *Fuel*, vol. 320, p. 123971, Jul. 2022, doi: 10.1016/j.fuel.2022.123971.

[54] P. Chen *et al.*, “Communications in Computer and Information Science 955 Commenced Publication in 2007 Founding and Former Series Editors: Editorial Board.” [Online]. Available: http://www.springer.com/series/7899

[55] C. E. Rasmussen, “Gaussian Processes in Machine Learning,” 2004, pp. 63–71. doi: 10.1007/978-3-540-28650-9_4.

[56] T. G. Dietterich, “Ensemble Methods in Machine Learning,” 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[57] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, Aug. 1999, doi: 10.1613/jair.614.

[58] Y. Wu and J. Feng, “Development and Application of Artificial Neural Network,” *Wirel Pers Commun*, vol. 102, no. 2, pp. 1645–1656, Sep. 2018, doi: 10.1007/s11277-017-5224-x.

[59] P. Kumar *et al.*, “Review of Nitrogen Compounds Prediction in Water Bodies Using Artificial Neural Networks and Other Models,” *Sustainability*, vol. 12, no. 11, p. 4359, May 2020, doi: 10.3390/su12114359.

[60] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” Nov. 2018.

[61] Arya, “What is Cross-Validation in Machine Learning?”

[62] J.R. Parker, *Algorithms for Image Processing and Computer Vision*, Second Edition. Indiana: Wiley Publishing, 2011.

[63] D. Berrar, “Cross-Validation,” in *Reference Module in Life Sciences*, Elsevier, 2024. doi: 10.1016/B978-0-323-95502-7.00032-4.

[64] R. Maredia, “Analysis of Google Play Store Data set and predict the popularity of an app on Google Play Store.” [Online]. Available: https://www.researchgate.net/publication/343769728

[65] G. Naidu, T. Zuva, and E. M. Sibanda, “A Review of Evaluation Metrics in Machine Learning Algorithms,” 2023, pp. 15–25. doi: 10.1007/978-3-031-35314-7_2.

[66] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[67] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geosci Model Dev*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[68] L. Ricci and R. Martínez, “Adjusted -type measures for Tweedie models,” *Comput Stat Data Anal*, vol. 52, no. 3, pp. 1650–1660, Jan. 2008, doi: 10.1016/j.csda.2007.05.017.

[69] H. M. Torun, M. Swaminathan, A. Kavungal Davis, and M. L. F. Bellaredj, “A Global Bayesian Optimization Algorithm and Its Application to

Integrated System Design," *IEEE Trans Very Large Scale Integr VLSI Syst*, vol. 26, no. 4, pp. 792–802, Apr. 2018, doi: 10.1109/TVLSI.2017.2784783.

[70] N. Loka, I. Couckuyt, F. Garbuglia, D. Spina, I. Van Nieuwenhuyse, and T. Dhaene, "Bi-objective Bayesian optimization of engineering problems with cheap and expensive cost functions," *Eng Comput*, vol. 39, no. 3, pp. 1923–1933, Jun. 2023, doi: 10.1007/s00366-021-01573-7.

[71] 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. 1, p. 22584, Dec. 2022, doi: 10.1038/s41598-022-27054-5.

PUBLICATIONS FROM THE THESIS

Conference Paper

’Eid, F. ’Sadikoglu, H. ’Akin I. M. (2023). Determination of pH Value Using Conventional pH Strips by Image Processing . In Prof. Dr. Muhittin DOĞAN (Ed.), *6th International Eurasian Conference on Biological and Chemical Sciences (EurasianBioChem 2023)* (pp. 325–325). www.eurasianbiochem.org .