International Journal of Nanoelectronics and Materials
Volume 14 (Special Issue) December 2021 [353-363]

Prediction Model for Spectroscopy Using Python Programming

A A M Ismail[1], N Ali[1,2]*, M S Amirul[2], R Endut[2] and S A Aljunid[2]

[1]Faculty of Electronic Engineering Technology (FTKEN), Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia.
[2]Advanced Communication Engineering, Centre of Excellence (CoE), Universiti Malaysia Perlis, 01000, Kangar, Perlis, Malaysia.

ABSTRACT

This paper is motivated by searching for the perfect pattern for the spectroscopy spectra using artificial neural networks (ANN) using python programming coding. The pattern from the spectroscopy is based on the absorption and emission of light and other radiation by materials in relation to the wavelength dependence of these processes. Spectral equipment such as spectrometers, spectral analysers, spectrographs, or spectrophotometers is utilised to determine spectrum values. The problem in this spectroscopy is to identify the sample or analyte, which can be solved by a prediction model for spectroscopy using Python. These problems occur when finding the best algorithm of pre-processing techniques that can predict any model accurately into an understandable format for prediction models. Various types of pre-processing techniques have been used, such as Multiplicative Scatter Correction (MSC), Inverse MSC, Extended MSC (EMSC), Extended Inverse MSC, de-trending, Standard Normal Variate (SNV) and normalisation in order to get a better r2 value. In this project, we find the r2 and the root mean square error (RMSE) to evaluate the prediction values and the actual values. First, choosing pre-processing techniques and then finding the best statistical method for constructing predictive models that produce high accuracy. We used ANN in this project as a prediction model. Based on the results, we managed to achieve our objective, which is that the prediction model has more than 90% of accuracy. Furthermore, the results show that our prediction model has 1.0 accuracy at 100 Epoch with a 0.3 learning rate. Finally, we can conclude that our prediction model can be used to predict the spectroscopy-based data format.

Keywords: artificial neural network, prediction model, python, spectroscopy

1. INTRODUCTION

Spectroscopy is the study of the absorption and emission of light and other radiation by matter related to the radiation wavelength dependence of these processes [1]. Spectral equipment such as spectrometers, spectral analysers, spectrographs, or spectrophotometers is used to calculate spectral values. This equipment is used to characterise and analyse research like dye-sensitised solar cells (DSSC) [2,3] and light-emitting diodes (LED) [4]. In our daily life, the observation of colour can be related to spectroscopy. For example, the standard emission frequencies of neon and other noble gases. To trigger these emissions, neon lamps use the collision of electrons with gas. Predictive models in Python are beneficial for predicting future results and estimating parameters that are hard or not practical to calculate [5]. For example, researchers can use predictive models to forecast crop yields based on rainfall and temperature or decide if patients with specific characteristics are more likely to respond. However, the researcher most uses linear regression for predictive modelling because of straightforward supervised machine learning algorithms.

*norshamsuri@unimap.edu.my

This project aims to create a new combination of pre-processing algorithm techniques to transform raw data into an understandable format for the prediction model. Pre-processing algorithm techniques are selected and tested and analysed consequently. The best algorithm pre-processing techniques are determined by observing the value of the determination (R2) coefficient, and root mean square error (RMSE) that could be nearest to 1. Then, the prediction data be measured by using suitable equipment to measure the spectral data of the model that is to be predicted. An example of the prediction model could be any matter that can absorb and immerse the light in the same radiation. Python is used because of its capability for analysing statistical data, and it is open-source software.

There are several things that we want to accomplish during this experiment. Firstly, to design suitable algorithm pre-processing techniques that can predict any model to the most accurate understandable format for prediction models. Next, we want to propose the best statistical method for constructing predictive models that is dependent on many factors and is highly collinear. Finally, we want to use Python to analyse the data collected from determination (R2), and root mean square error (RMSE) that could be nearest to 1. This article is to provide information and details on the prediction model for spectroscopy using Python programming. The structure of this article is generally divided into four parts. Part 1 is about the introduction of spectroscopy and prediction models based on Python. This topic had explained the aim and motivation, and objectives to conduct this project. The next topic is the literature review. It is about the study from research related to this project. The research is mainly done from the website, and some information is also obtained from the paper, journal, etc. In part 3, it is discussing the project's methodology. This chapter contains several parts: the introduction, flowchart for the prediction model for spectroscopy using Python. In part 4, it discussed the results and conclusion based on this project.

In almost all scientific areas of science and technology, spectroscopic techniques have been applied. For example, magnetic Resonance Imaging (MRI) is the term used in a medical technique. It images the body's internal soft tissue with unparalleled precision using the radio-frequency spectroscopy of nuclei in a magnetic spectrum [6]. As a result, spectroscopy now covers a sizable fraction of the electromagnetic spectrum.

1.1 Pre-processing Technique

Pre-processing is a technique of data mining that processes unprocessed data into a suitable form format. Unprocessed data, which is real-world data, is consistently inaccurate and cannot be sent through a model. That would cause specific errors for the prediction model. That is why the pre-processing technique is needed to pre-process data before sending it through a model. The steps for data pre-processing are shown in Figure 1.

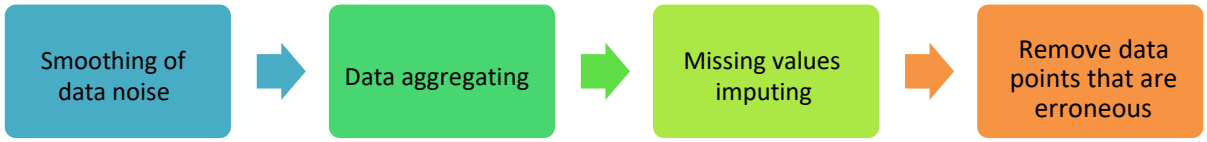


Figure 1. Steps in data pre-processing

In previous studies, they used five pre-processing techniques that have been used for estimation of Harumanis (Mangifera indica L.) sweetness using Near-Infrared (NIR) Spectroscopy [7]. Therefore, unit vector normalisation (UVN), Multiplicative scatter correction using common amplification (MSCCA), multiplicative scatter correction using common offset (MSCCO), multiplicative scatter correction (MSC) and mean normalisation (MN) technique is selected to treat the spectral data in this study because of their optimum performance in the PLSR model testing. To decide the best pre-processing strategy, a comparison of these pre-processing

methods was carried out, and the results were concluded by measuring the determination coefficient (R2) and the Root Mean Square Error (RMSE) [8].

1.1.1 Determination of coefficient (R2)

The determination coefficient R2 is the proportion of the variability in the data captured by the model evaluated. R2 is measured on a scale from 0 and 1, where 0 is no correlation is implied, and 1 is a perfect correlation between the variables. It should be noted that the bias between the expected and the actual value is not considered in R2. R2 can be calculated by using equation 1:

$$R^2 = \left(\frac{\boldsymbol{n}(\sum \boldsymbol{xy}) \; - \; (\sum \boldsymbol{x})\,(\sum \boldsymbol{y})}{\sqrt{[\boldsymbol{n}\sum \boldsymbol{x^2} - (\sum \boldsymbol{x})^2\,][\boldsymbol{n}\sum \boldsymbol{y^2} - (\sum \boldsymbol{y})^2]}} \right)^{\mathbf{2}} \tag{1}$$

1.1.2 Root Mean Square Error (RMSE)

To find the prediction error, RMSE as in Equation (2) is used. The prediction of error is used when we want to know how far the value can be from the regression line's data point. It is common and suitable to verify the experimental results to use Root Mean Square Error in forecasting, climatology, and regression analysis [9].

$$RMSE = \sqrt{\overline{(f - o)^2}} \tag{2}$$

Where:

f = forecasts (expected values or unknown results),

o = observed values (known results).

1.2 Prediction Model

To recognise patterns and trends within historical data, a predictive model uses statistical techniques. To predict those results or behaviours, predictive analytics algorithms are then implemented. Although this is dependent on historical data, fresh and near-real-time data must be fed into the predictive model so that the predictive model consistently learns and continuously increases the accuracy of its forecasts [10]. We can develop a broad range of data models such as regression, classification, clustering, decision tree and time series models with the best predictive analysis tools [11]. Predictive modelling is often achieved using curve and surface fitting, time series regression, or machine learning methods. Regardless of the approach used, the method of constructing a model prediction through methods is similar. The steps for constructing a prediction model are shown in Figure 2:

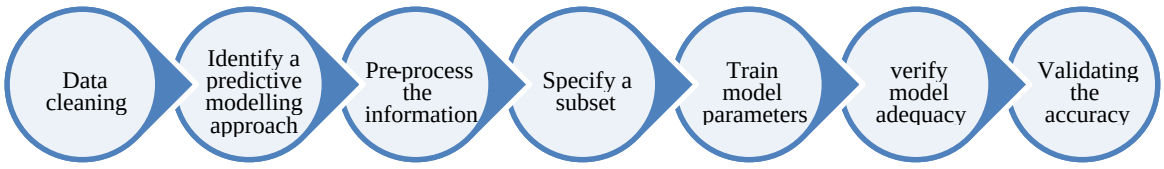


Figure 2. The steps of constructing a predictive model

1.2.1 Artificial Neural Network (ANN).

An artificial neural network (ANN) is the piece of a computing system designed to simulate the way the human brain analyses and processes information [12]. It is the basis of artificial intelligence (AI) and solves problems that, by human or mathematical standards, would prove impossible or challenging. ANNs have self-learning capabilities that allow them, as more data becomes available, to deliver better results. The ANN-based approach can be used to model dynamic interactions or to identify similarities in data between inputs and outputs. ANN can be viewed as a mathematical model or computational model that is inspired by the structure or functional aspects of biological neural networks [13]. Neural networks are designed to extract existing patterns from noisy data. The procedure involves training a network (training phase) with a large sample of representative data, after which one exposes the network to data not included in the training set (validation or prediction phase) with the aim of predicting the new outcomes [14]. Figure 3 is an example of an artificial neural network (ANN) structure.

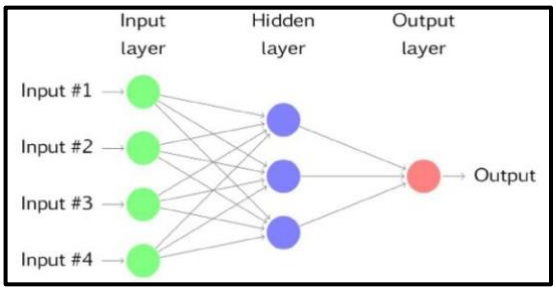


Figure 3. A neural network with four inputs and one hidden layer

2. MATERIAL AND METHODS

To design the suitable algorithm pre-processing techniques that can predict any model to most accurately into an understandable format for prediction models, data mining problems are usually broken down into several tasks that can be used to group techniques and application areas [15]. There are four steps involved in the design of the pre-processing techniques. Firstly, smoothing of noisy data. Biological recordings can be extremely noisy, so it is always important to filter the data. Data is likely to be obtained concurrently by various recording devices, probably at different temporal or spatial resolutions, and need to be aggregated into the same tables or matrices, possibly with suitable subsampling. Next is imputing missing values. In analysis scripts, taking the time to perform the proper error handling for missing values or Nans ("Not-a-Number") can save hours of debugging further down the analysis pipeline. \

Next is to propose the best statistical method for constructing predictive models that depend on many factors and is highly collinear. The aim of predictive modelling is to minimise the difference between the values that are expected and actual [16]. A model representation is made up of a set of parameters organised in structure (attributes, operators, and constants). Predictive modelling is the method of tuning or training the model parameters using a data mining algorithm to conform as much as possible to a collection of instances of the definition. Python is used to analyse the data collected from determination (R2) and Root Mean Square Error (RMSE) that could be nearest to 1. All Python libraries concentrate on creating one thing that makes data analysis more straightforward, more open, and thorough. The process flow of this project is shown in Figure 4.

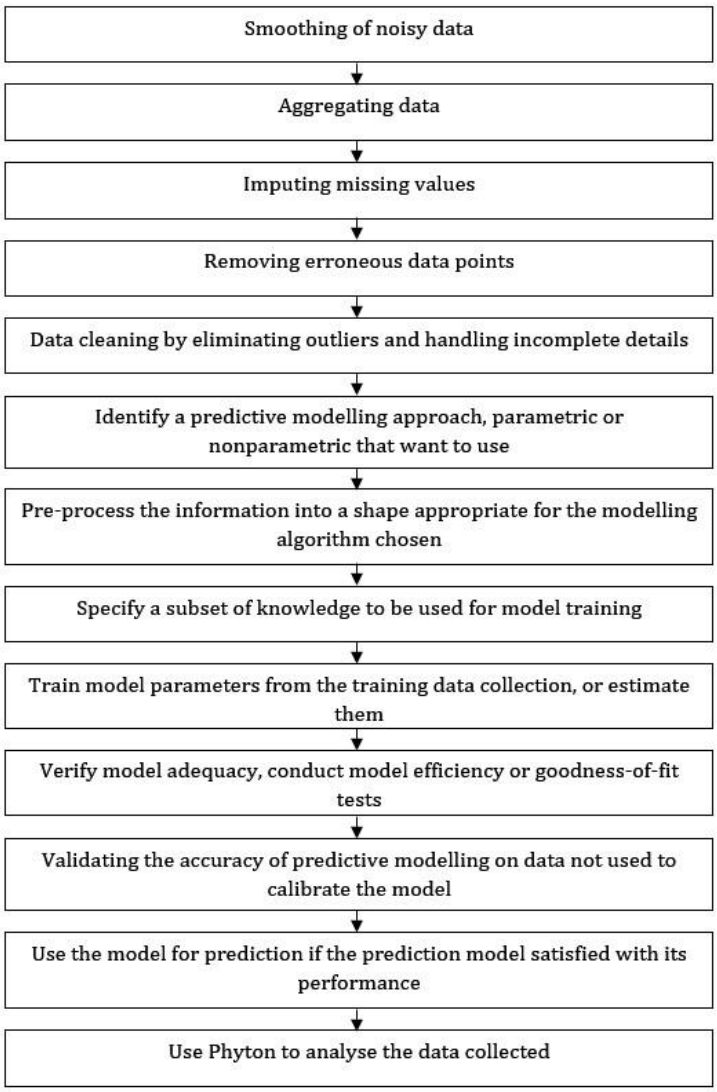


Figure 4. The process flow of constructing the prediction model for Spectroscopy

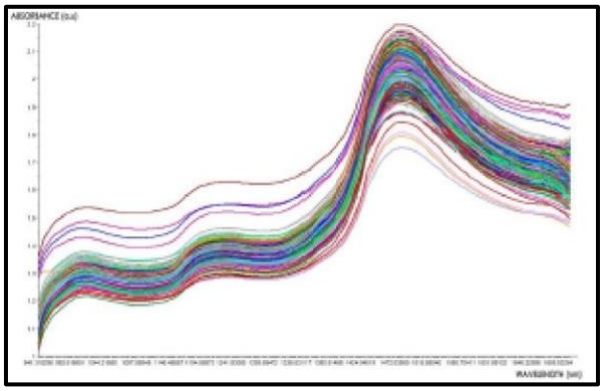
3. RESULTS AND DISCUSSION

In results, we have the results of training on data sets that can be used as predictive modelling. These results are obtained by starting on data exploration and undergo the process of data cleaning. Then, the cleaned data can be analysed using suitable modelling such as linear regression. Finally, we get the performance analysis on the ability to predict the target object.

3.1 Data Sets

From the data sets below, there are values P (Phosphorus) from the soil. These values were obtained by undergoing a dilution process called Hanna to find the amount of P (Phosphorus) in the soil. For this project, we would like to find the value of P that is obtained from the amount of the reflection that measures, using the spectrometer in figure 5 (a). Figure 5 (b) shows the raw data of the wavelength of the spectral wave and its reflection. This process is repeated by doubling the amount of dilution until the amount of P (Phosphorus) reaches the minimum amount of dilution, which is at 23.5 amount of P (Phosphorus). The amount that was measured

was between 61.8 until 23.5. Figure 5 (c) shows the CSV format that is used in Python programming.



(a)

P- value obtain on Hanna

P= 61.8		P= 50.3	
WaveLeng	Reflection	WaveLeng	Reflection
901.116	0	901.116	1.351623
905.026	1.735954	905.026	2.051413
908.932	1.756463	908.932	1.62683
912.834	1.77313	912.834	1.763722
918.029	1.650357	918.029	1.581789
921.921	1.861401	921.921	1.593723
925.808	1.738492	925.808	1.632982
929.691	1.656019	929.691	1.675865
933.57	1.588556	933.57	1.630445
937.445	1.682143	937.445	1.610397
941.315	1.606974	941.315	1.599031
945.182	1.606879	945.182	1.629281
949.043	1.613996	949.043	1.582286
952.901	1.616147	952.901	1.591348

(b)

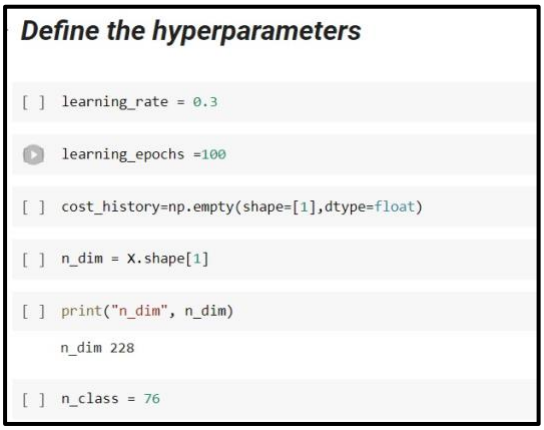
A	B	C	D	E	F	G	H	I
0	1.73595	1.75646	1.77313	1.65036	1.8614	1.73849	1.65602	1.58856
1.35162	2.05141	1.62683	1.76372	1.58179	1.59372	1.63298	1.67587	1.63045
0	1.75629	2.28004	1.59388	1.88868	1.80152	1.66636	1.68513	1.68347
1.73962	1.31534	1.89346	2.57741	1.83914	1.69099	1.67268	1.66647	1.59508
2.04892	1.87925	1.49956	1.63801	1.6867	1.59228	1.72393	1.58916	1.63499
1.84064	1.31212	1.85495	1.56644	1.52428	1.633	1.66096	1.618	1.61618
0	1.92065	2.43897	1.48825	2.0393	1.65851	1.68778	1.60473	1.61618
3.46389	2.40213	1.71641	1.50416	2.13205	1.64313	1.53561	1.7488	1.64989
0	2.12825	1.83691	1.7455	1.57219	1.74959	1.68998	1.64919	1.62713
1.8355	1.4523	2.23993	1.49655	1.60233	1.59037	1.62493	1.60085	1.6314
3.2878	0	2.02711	3.21088	1.72456	1.56711	1.70161	1.728	1.61572
2.68574	1.52445	1.47255	1.97788	1.64583	1.76933	1.6869	1.62639	1.67787
0	0	0	1.46737	1.83217	1.62724	1.65116	1.69461	1.63283
1.33517	1.64418	1.51734	1.96784	1.84861	1.57212	1.6174	1.69154	1.63619
0	0	1.18369	1.81643	1.61263	1.63142	1.50688	1.61287	1.6949
2.48617	2.09457	2.0596	1.94137	2.03376	1.63565	1.7505	1.65822	1.66276
1.93241	1.59023	1.92631	1.54249	1.39864	1.76646	1.68646	1.60251	1.60618
1.28348	0	1.46071	1.54249	1.8498	1.64911	1.60962	1.63764	1.59442
3.16286	1.90361	1.96863	1.39398	1.35968	1.50596	1.48597	1.54799	1.54544
0	1.89126	1.7824	1.73959	1.53636	1.55417	1.51359	1.71634	1.57768
1.25038	1.11447	1.39574	1.49238	1.62964	1.74275	1.63221	1.7332	1.56646

(c)

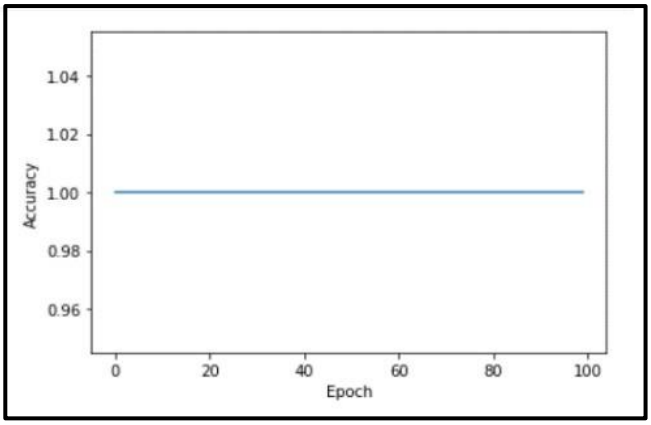
Figure 5. (a) The Spectral plot,(b) The raw data and (c) The CSV format

3.2 RMSE and Prediction performance

From this project, we are able to find the root mean square error (RMSE) from the data set that we used. The prediction model that has been constructed managed to get high accuracy of predicted values which is the value of P (Phosphorus). Figure 6 (a) is the hyperparameters that have been set to the prediction model. The learning rate is set at 0.3, the Epoch which indicates the number of passes of the entire training dataset that has been completed and is set at 100 and the number of classes is 76. We used four hidden layers in this model. Figure 6 (b) shows the prediction accuracy against the Epoch of the prediction model. Finally, Figure 6 (c) shows the RMSE values and the training accuracy of the model. As shown in Figure 6(c) we manage to get almost perfect analysis with our data set that we compare to the value measured by the lab equipment [17].



(a)



(b)

```
    ...
   [4.54116845e+12 4.54117619e+12 4.54118753e+12 ... 4.54120437e+12
    4.54114927e+12 4.54118441e+12]
   [4.54116845e+12 4.54117619e+12 4.54118753e+12 ... 4.54120437e+12
    4.54114927e+12 4.54118441e+12]
   [4.54116845e+12 4.54117619e+12 4.54118753e+12 ... 4.54120437e+12
    4.54114927e+12 4.54118441e+12]]
epoch:  99  -  cost:  3.3530905170654255  - MSE:  1.5352097804477933e+26 - Train Accuracy:  1.0
```

(c)

Figure 6. (a)The hyperparameters of ANN, (b) The prediction accuracy against the Epoch and (c) The training accuracy and RMSE values

4. CONCLUSION

In conclusion, spectroscopy is the study of the absorption and emission of light and other radiation by matter related to the radiation wavelength dependence of these processes. These can be measured by using specific equipment such as spectral analysers or spectrometers. The predictive model for this project had undergone a pre-processing of data and findings the best prediction model based on the datasets that were obtained. We also managed to achieve our objective that the prediction model can predict more than 90% accuracy. There are also some improvements that can be made or in future such as we can use the prediction algorithm in the microcontroller for prediction purposes.

ACKNOWLEDGEMENTS

We would like to acknowledge the support from the Centre of Excellence Advance Communication Engineering (ACE) Optics, especially the Optic group team, for providing the spectral datasets. The research has been carried out under Malaysia Technical University Network (MTUN) 2019 Matching Research Grant 9028-00004 provided by the Ministry of Higher Education of Malaysia (MOHE).

REFERENCES

[1] Paldus, B. A., and R. N. Zare. "Absorption spectroscopies: from early beginnings to cavity-ringdown spectroscopy," in Cavity-ringdown spectroscopy: an ultratrace-absorption measurement technique, Busch, Kenneth W., and Marianna A. Busch, eds.: American Chemical Society, (1999) pp. 49-70.

[2] Jamalullail, N., Mohamad, I. S., Norizan, M. N., Baharum, N. A. and Mahmed, N., "Short review: Natural pigments photosensitiser for dye-sensitised solar cell (DSSC)," in IEEE 15th Student Conference on Research and Development (SCOReD) (IEEE), (2017) pp. 344–9.

[3] Jamalullail, N., Mohamad, I. S., Norizan, M. N., Mahmed, N., & Taib, B. N., "Recent improvements on TiO2 and ZnO nanostructure photoanode for dye sensitized solar cells: A brief review," EPJ Web of Conferences Vol. **162** (2017) pp. 01045.

[4] Abdul Rais, S. A., Hassan, Z., Abu Bakar, A. S., Abd Rahman, M. N., Yusuf, Y., Md Taib, M. I., Sulaiman, A. F., Hussin, H. N., Ahmad, M. F., Norizan, M. N., Nagai, K., Akimoto, Y. and Shoji, D., "Effect of indium pre-flow on wavelength shift and crystal structure of deep green light emitting diodes," Opt. Mater. Express, **11**, (2021) pp. 926.

[5] Doupe, P., Faghmous, J. and Basu, S., "Machine learning for health services researchers," Value in Health, vol. **22**, issue 7, (2019) pp. 808-815.

[6] BRADBURY, E.M., RADDA, G.K. and ALLEN, P.S., "Nuclear magnetic resonance techniques in medicine," Annals of internal medicine, vol. **98**, issue 4, (1983) pp. 514-529.

[7] Amirul, M.S., Endut, R., Rashidi, C.B.M., Aljunid, S.A., Ali, N., Laili, M.H., Laili, A.R. and Ismail, M.N.M., "Estimation of Harumanis (Mangifera indica L.) Sweetness using Near-Infrared (NIR) Spectroscopy," IOP Conference Series: Materials Science and Engineering vol. **767**, No. 1 (2020) pp. 012070.

[8] Sabri, M.S.A., Endut, R., Rashidi, C.B.M., Laili, A.R., Aljunid, S.A. and Ali, N., "Analysis of Near-infrared (NIR) spectroscopy for chlorophyll prediction in oil palm leaves," Bulletin of Electrical Engineering and Informatics, vol. **8**, issue 2, (2019) pp. 506-513.

[9] Das, B., Nair, B., Reddy, V.K. and Venkatesh, P., "Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India," International journal of biometeorology, vol. **62**, issue 10, (2018) pp. 1809-1822.

[10] Nair, R., Hoang, T.L., Laumanns, M., Chen, B., Cogill, R., Szabó, J. and Walter, T., "An ensemble prediction model for train delays," Transportation Research Part C: Emerging Technologies vol. **104** (2019) pp. 196-209.

[11] Lai, R.K., Fan, C.Y., Huang, W.H. and Chang, P.C., "Evolving and clustering fuzzy decision tree for financial time series data forecasting," Expert Systems with Applications vol. **36**, issue 2, (2009) pp. 3761-3773.

[12] Agatonovic-Kustrin, S. and Beresford, R., "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," Journal of pharmaceutical and biomedical analysis, vol. **22**, issue 5 (2000) pp.717-727.

[13] Braik, M., Al-Zoubi, H. and Al-Hiary, H., "Artificial neural networks training via bio-inspired optimisation algorithms: modelling industrial winding process, case study," Soft Computing, vol. **25**, issue 6, (2021) pp. 4545-4569.

[14] Hassan, A., Baksh, M.S.N., Shaharoun, A.M. and Jamaluddin, H., "Improved SPC chart pattern recognition using statistical features," International Journal of Production Research, vol. **41**, issue 7 (2003) pp.1587-1603.

[15] Gupta, Gopal K., "Introduction to data mining with case studies," PHI Learning Pvt. Ltd., (2014).

[16] Bodin, J., Ackerer, P., Boisson, A., Bourbiaux, B., Bruel, D., de Dreuzy, J.R., Delay, F., Porel, G. and Pourpak, H., "Predictive modelling of hydraulic head responses to dipole flow experiments in a fractured/karstified limestone aquifer: Insights from a comparison of five modelling approaches to real-field experiments," Journal of Hydrology vol. **454**, (2012) pp. 82-100.

[17] Amirul, M.S., Endut, R., Rashidi, C.B.M., Aljunid, S.A., Ali, N., Laili, M.H., Laili, A.R. and Ismail, M.N.M., "Nitrate (NO3-) prediction in soil analysis using near-infrared (NIR) spectroscopy," AIP Conference Proceedings vol. **2203**, No. 1 (2020) pp. 020043.