Prediction of Student's Academic Performance during Online Learning Based on Regression in Support Vector Machine

Nor Ain Maisarah Samsudin, Shazlyn Milleana Shaharudin, Nurul Ainina Filza Sulaiman, Shuhaida Ismail, Nur Syarafina Mohamed, and Nor Hafizah Md Husin

***Abstract*—Since the Movement Control Order (MCO) was adopted, all the universities have implemented and modified the principle of online learning and teaching in consequence of Covid-19. This situation has relatively affected the students' academic performance. Therefore, this paper employs the regression method in Support Vector Machine (SVM) to investigate the prediction of students' academic performance in online learning during the Covid-19 pandemic. The data was collected from undergraduate students of the Department of Mathematics, Faculty of Science and Mathematics, Sultan Idris Education University (UPSI). Students' Cumulative Grade Point Average (CGPA) during online learning indicates their academic performance. The algorithm of Support Vector Machine (SVM) as a machine learning was employed to construct a prediction model of students' academic performance. , Two parameters, namely C (cost) and epsilon of the Support Vector Machine (SVM) algorithm should be identified first prior to further analysis. The best parameter C (cost) and epsilon in SVM regression are 4 and 0.8. The parameters then were used for four kernels, i.e., radial basis function kernel, linear kernel, polynomial kernel, and sigmoid kernel. from the findings, the finest type of kernel is the radial basis function kernel, with the lowest support vector value and the lowest Root Mean Square Error (RMSE) which are 27 and 0.2557. Based on the research, the results show that the pattern of prediction of students' academic performance is similar to the current CGPA. Therefore, Support Vector Machine regression can predict students' academic performance.**

***Index Terms*—Support vector machine, regression, epsilon, cost, linear kernel, polynomial kernel, sigmoid kernel radial basis function kernel.**

I. Introduction

A novel coronavirus from Wuhan province, China, shocked the world on 31st December 2019. The outbreak is the latest form of Coronavirus (Covid-19) that has now been discovered to cause serious illness and death [1]. The World Health Organization (WHO) pronounced it a 'global health emergency' in January 2020 when the Covid-19 epidemic claimed 170 lives in a short time [2]. Later in March 2020, the WHO announced it as a pandemic that has spread across the region and involved large populations of the whole world [3]. Malaysia is also susceptible to the COVID-19 epidemic. Malaysia initially had a controlled rate of positive COVID-19 cases, but with the arrival of foreigners, the number of cases surged [4].

On March 16, 2020, Tan Sri Muhyiddin Yassin, Malaysia's Prime Minister, announced the Movement Control Order (MCO) as means to control the spread of Covid-19 [5]. This MCO has had the unintended consequence of preventing schools and institutions of higher learning from operating during the MCO time. On May 27, 2020, the Ministry of Higher Education (MOHE) imposed full online teaching and learning programmes for students by December 31 [6]. Sultan Idris Education University (UPSI) implemented online learning and teaching according to MOHE's charge [7].

Students at universities are not an exception to online learning. However, the overall phenomenon of digital lectures exhibits a significant effect on university students, particularly in terms of academic achievement [8]. There is no evidence to support the argument that online learning time correlates with CGPA even though spending more time on online learning was anticipated to result in higher academic performance [9]. Although numerous studies explored the topic of online learning, acquiring studies that focused on the context of pandemics was challenging [10].

Support Vector Machine (SVM) is one of the supervised machine learning that could solve classification and regression problems [11]. Time series prediction in SVM can be seen as potential use of regression in machine learning [12]. In [13], the study is about forecasting the rate patterns of student graduation with Na ïve Bayes Classifier and Support Vector Machine, and the result shows that the Support Vector Machine is more efficient than the Na ïve Bayes Classifier, displaying 69.15% accuracy of the overall data. The study revealed that Support Vector Regression-Based Prediction of Students' School Performance in [14] SVR (Support Vector Machine –Regression) is an appropriate method to explore personality correlations since the forecasted performances mostly generated approximately 80% accuracy. It also means that the regression in SVM could project the students' school performance. In [15], the result shows that SVR (Support Vector Machine – regression) and Linear Regression (LR) models are employable for the implementation of the

Manuscript received May 12, 2022; revised July 18, 2022. This work was supported in part by the Fundamental Research Grants Scheme Vote No. 2019-0132-103-02 (FRGS/1/2019/STG06/UPSI/02/4) granted by the Malaysian Ministry of Education.

N. A. M. Samsudin, S. M. Shaharudin, N. A. F. Sulaiman, and N. H. M. Husin is with the Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia (corresponding author: Shazlyn Milleana Shaharudin; e-mail: norainmaisarah28@gmail.com, shazlyn@fsmt.upsi.edu.my, aininafilza@gmail.com, hafizah.husin@fsmt.upsi.edu.my).

S. Ismail is with Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, 84600 Panchor, Johor, Malaysia (e-mail: shuhaida@uthm.edu.my).

N. S. Mohamed is with Department of Mathematical Sciences, Universiti Teknologi Malaysia, 81300 Skudai, Johor, Malaysia (e-mail: nursyarafina@utm.my).

doi: 10.18178/ijiet.2022.12.12.1768

university's Student Performance Prediction System.

This study aims to generate a projection of the academic performance of undergraduate students from the Department of Mathematics, UPSI during online learning by using regression in Support Vector Machine (SVM). Researchers found that the findings of this study provided many benefits and contributions, including the possibility of helping university administrators develop plans to increase efficiency in their institutions. Moreover, the present study provides a prediction model that can be used to evaluate students, which can be extremely useful with resource selection as part of intervention programs. This study focuses on the field of education that has to do with predicting academic achievement using SVM, a machine learning approach. Additionally, the study lays out guidelines for selecting appropriate parameters to apply with data on education, such as academic performance.

II. Methodology

To predict students' academic performance during online learning, the propositioned modelling of students' academic performance is illustrated in Fig. 1.

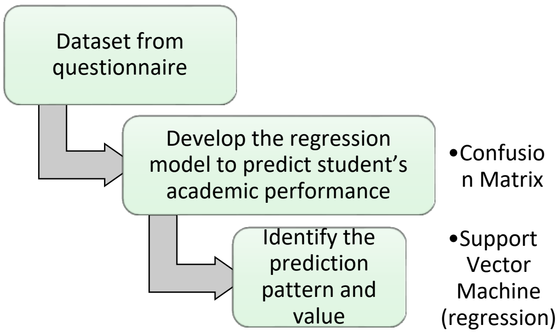
Dataset from questionnaire
Develop the regression model to predict student's academic performance
•Confusio n Matrix
Identify the prediction pattern and value
•Support Vector Machine (regression)

Fig. 1. Flowchart of prediction students' academic performance.

A. Data

For this study, the dataset employed was collected from undergraduate students at Sultan Idris Education University (UPSI). The data were gathered through a questionnaire and was issued via an online platform. It referred to the latest Cumulative Grade Point Average (CGPA). The responses to the questionnaires were from undergraduate students from the Faculty of Science and Mathematics, UPSI who were in semesters 3 to 7. Based on the data 82.5% were female and 17.5% were male. 72.5% were from the mathematics department, 5.7% from the Biology department, 12.2% from the Chemistry department, 3.1% from the Physics department, and 6.6% from the Science department, with a total of 225 student respondents.

B. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a highly efficient machine learning tool that was recommended by [16] and became more attractive to machine learning researchers and communities. The algorithm of SVM was evidently effective in regression and classification methods.

SVM regression performs linear regression in the high-dimensional feature space by employing ε^- insensitivity loss and, concurrently attempts to decrease model complexity by minimizing $||w||^2$. It may be defined through the introduction of slack variables ξi and ξi^*, where $i = 1, \dots, n$, could calculate the deviation of the training sample outside the ε^- sensitive zone.

$$\phi(w, \xi) = \frac{1}{2}\|w\| + C\sum_{i=1}^{n}(\xi i + \xi i^*) \tag{1}$$

$$\text{Min}\begin{cases} y_i - f(x_i, w) \le \varepsilon + \xi i^* \\ f(x_i, w) - y_i \le \varepsilon + \xi i \\ \xi i, \xi i^* \ge 0, i = 1, \dots, n \end{cases} \tag{2}$$

This optimization problem could become the Lagrangian dual problem, and the solution can be denoted as,

$$\text{Max } W(\alpha, \alpha^*) = -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i{}^*) + \sum_{i=1}^{n}(\alpha_i - \alpha_i{}^*) - \frac{1}{2}\sum_{ij=1}^{n}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)\langle x_i, x_j\rangle$$

Subject to

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i{}^*) = 0 \qquad \alpha_i{}^*, \alpha_i \,\epsilon\, [0, C], i = 1,2,3, \dots, n \tag{3}$$

In dual problem, kernel functions $K\langle x_i, x_j\rangle$ is used to substitute $\langle x_i, x_j\rangle$. The desired regression function is then:

$$f(x) = \sum_{i=1}^{n}(\alpha_i{}^* - \alpha_i)K(x, x_i) + b \tag{4}$$

Some different kernels were employed in SVM regression, but the standard kernels employed were linear, polynomial, radial basis function, and sigmoid. The most well-known and qualified kernel is known to be the radial basis function with parameter γ.

$$K(x, x') = \begin{cases} x^T.x' & linear \\ (x^T.x' + 1)^d & polynomial \\ \exp(-\gamma||x - x'||^2) & RBF \\ \tanh(\gamma x.x' + C) & sigmoid \end{cases} \tag{5}$$

To evaluate the result from the SVM regression model, equation (6) was used to measure the error rate:

$$RMSError = \sqrt{1 - r^2}SD_y \tag{6}$$

where SD_y is the standard deviation of Y. Root Mean Square (RMSE) is the standard deviation of the residual (prediction errors).

III. Result and Discussion

The x-axis and the y-axis from Fig. 2 represent the students and the GPAs. Fig. 2 displays the result of GPA students before and during pandemic outbreaks by employing two class types. The blue line denotes the students' GPA when face-to-face classes were conducted, and the orange line denotes the student's GPA when online classes were conducted. From Fig. 2, the pattern demonstrates how students mainly obtained exceptional results when participating in online classes compared to face-to-face classes. It is due to the orange line being mostly above the blue line, indicating that most of the students scored better during online classes than the face-to-face classes. Hence, employing a prediction model could estimate the student's GPA performance in the following semester should online classes be maintained.

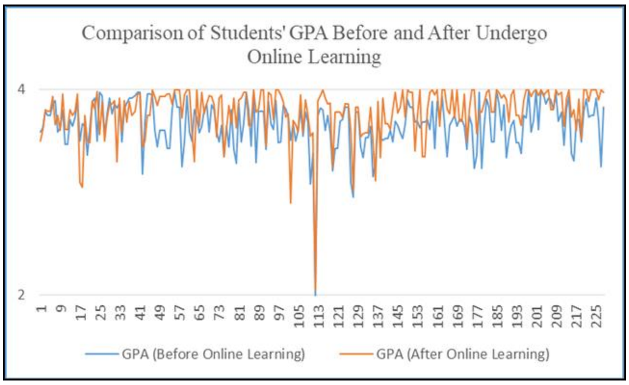

Fig. 2. GPA of academic performance students before and during COVID-19 outbreak.

The SVM regression model was employed to project students' academic performance. Consequently, it is highly vital to determine the most fitting parameters in SVM regression. Choosing the best ones could enhance the prediction accuracy while the process of selection was named by turning the parameter in SVM. Table I illustrates the turning parameter process and the most fitting parameter selected according to the minimum error. The parameter C values chosen were 4, 8, 16, 32, 64, and 128. In addition, the chosen epsilon values are 0, 0.2, 0.4, 0.6, 0.8, and 1 for each value of C, respectively. Therefore, from Table I, the minimum error is found to be 0.03825957.

TABLE I: THE RESULT OF TURNING PARAMETER PROCESS

	Cost	Epsilon	Error	Dispersion
1	4	0.2	0.041470	0.024125
2	8	0.2	0.042042	0.024588
3	16	0.2	0.042679	0.024649
4	32	0.2	0.043078	0.024504
5	64	0.2	0.043448	0.024481
6	128	0.2	0.043402	0.024276
7	4	0.4	0.040517	0.022203
8	8	0.4	0.040759	0.022755
9	16	0.4	0.040716	0.022979
10	32	0.4	0.040589	0.022957
11	64	0.4	0.040318	0.022759
12	128	0.4	0.040418	0.022778
13	4	0.6	0.039169	0.020091
14	8	0.6	0.038573	0.019832
15	16	0.6	0.038651	0.019888
16	32	0.6	0.038791	0.020202
17	64	0.6	0.038929	0.02017
18	128	0.6	0.039085	0.02033
19	4	0.8	0.038260	0.017541
20	8	0.8	0.038264	0.017537
21	16	0.8	0.038375	0.017569
22	32	0.8	0.038387	0.01758
23	64	0.8	0.038413	0.017599
24	128	0.8	0.038271	0.017499
25	4	1	0.041135	0.016689
26	8	1	0.041062	0.016982
27	16	1	0.041160	0.017106
28	32	1	0.041533	0.017369
29	64	1	0.041476	0.017118
30	128	1	0.041659	0.016993

However, according to Table I, there was a minor difference in the error patterns. Therefore, the results of measuring the performance of turning parameters are illustrated graphically to determine the most fitting parameter pairs. Fig. 3 shows the relationship between cost and epsilon and the resulting error. The error was calculated via the colour reference scale on the right side of Fig. 3. From the result, Fig. 3 shows that as epsilon becomes 0.6, the darker it becomes, and as it approaches a bigger cost value, it becomes darker than the rest of the values. Therefore, it can be concluded that the optimal parameter pair is to select a C value of 128 and an epsilon value of 0.63. With the selected parameter pair the SVM regression will produce the best accuracy. To achieve the best results using SVM in this study, it is also important to emphasize the selection of kernels. Radial basis function, sigmoid, polynomial, and linear kernel were used to derive the kernel parameters in this study. Each kernel function possesses its specific parameter that needs optimisation for achieving the most valuable result performance [17].

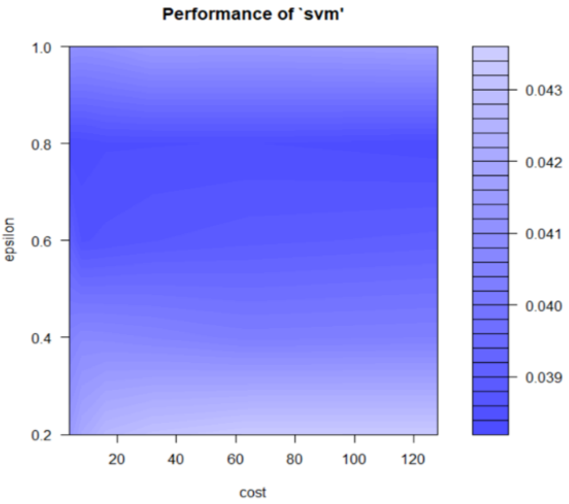

Fig. 3. Performance of turning parameter.

TABLE II: THE RESULT OF SUPPORT VECTOR MACHINE BY REGRESSION

Type of kernel	Parameter		Result	
	C	Epsilon	No. of support vectors	RMSE
Linear	4	0.8	29	0.2661
Radial Basis Function (RBF)	4	0.8	27	0.2557
Polynomial	4	0.8	29	0.2649
Sigmoid	4	0.8	110	4.3823

In Table II, SVM results were achieved by determining C, the total of support vectors, and the value of RMSE. The total of support vectors represents the data that approaches or is distant from the hyperplane during regression. The appropriate number of support vectors is 27 which represents regression as a minimum value. Table II shows that the radial basis function (RBF) kernel has the lowest RMSE around 0.2555, while the sigmoid kernel has the highest RMSE around 4.3823.

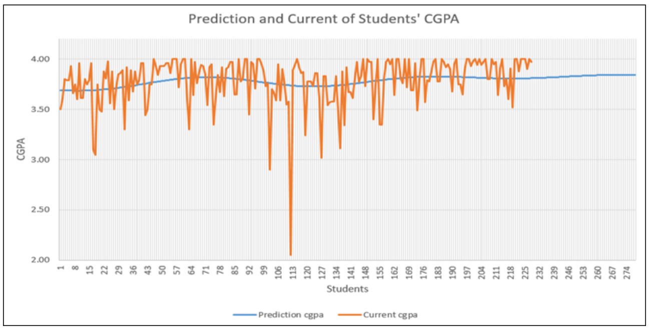

Fig. 4. Current and forecasted CGPA during online learning.

From Fig. 4, the x-axis and the y-axis represent the students and the CGPAs. Fig. 4 shows the result of current and forecasted students' CGPA during a pandemic outbreak. The orange line represents the current CGPA, and the blue line represents the forecasted CGPA. The result shows that the prediction by using SVM regression can follow the pattern of the current CGPA. It also can forecast almost another 50 students' CGPA. However, it is unable to fully cater to the variation since there was a lack of dataset, and more variables were needed.

IV. Conclusion

Identifying and predicting a student's academic performance is beneficial for various situations, more so in management, like determining scholarship programs, enrolling outstanding students, and helping to recognise those with less graduation possibility. In addition, this research also helped us to identify some important patterns of students' academic performance at the Sultan Idris Education University (UPSI). In this study, the researcher used the data of students' CGPA during online learning to predict students' academic performance by using SVM regression. From the findings through the employment of SVM regression, the most suitable parameters C and epsilon are 4 and 0.8. The radial basis function with the lowest number of support vectors which is 27 was chosen as the best type of kernel also resulting in the lowest value of RMSE which is 0.2557. The pattern of prediction of students' academic performance is similar to the current CGPA. Therefore, Support Vector Machine regression can predict students' academic performance. The limits ascertained in this study need an emphasis when utilising the predictions made by the SVM model to evaluate students' academic performance. In this study, we only used 228 samples and 2 feature data sets. However, because the data points of all two functions were inconsistent, it is impossible to collect a larger sample size of the test and training data sets. Therefore, further expansion of sample size and features will further increase its accuracy.

Conflict of Interest

The authors of this study proclaim no conflict of interest.

Author Contributions

Conceptualization, Shazlyn Milleana Shaharudin; Nor Hafizah Md Husin; methodology, Nurul Ain Maisara Samsudin; software, Nurul Ain Maisara Samsudin; validation, Shuhaida Ismail; writing—original draft preparation, Nurul Ain Maisara Samsudin; writing—review and editing, Nurul Ainina Filza Sulaiman, Shazlyn Milleana Shaharudin, Nur Syarafina Mohamed; project administration, Shazlyn Milleana Shaharudin. The authors have collectively perused and approved the issued manuscript.

References

[1] World Health Organization: WHO. (February 2020). A joint statement on tourism and COVID-19 - UNWTO and WHO call for responsibility and coordination. [Online]. Available: https://www.who.int/news/item/27-02-2020-a-joint-statement-on-tourism-and-covid-19---unwto-and-who-call-for-responsibility-and-coordination#:~:text=On%2030%20January%202020%2C%20the,a%20set%20of%20Temporary%20Recommendations

[2] World Health Organization. (January 2020). Pneumonia of unknown cause — China. [Online]. Available: https:/entity/csr/don/05-january-2020-pneumonia-of-unknown-cause-china/en/index.html

[3] Ministry of Health Malaysia. (2018). Timeline: WHO's COVID-19 response. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline?gclid=CjwKCAiA4rGCBhAQEiwAelVti5CodUzm1UcfIW9uEPqKVvUJVYm0XJHcuZ7cVbRtDKeAYYDWWk0ZthoCoD0QAvD_BwE#event-115

[4] A. Tang. (March 16, 2020). Malaysia announces movement control order after spike in Covid-19 cases (updated). *The Star Online*. [Online]. Available: https://www.thestar.com.my/news/nation/2020/03/16/malaysia-announces-restricted-movement-measure-after-spike-in-covid-19-cases

[5] (Mar. 16, 2020). Covid-19: Movement control order imposed with only essential sectors operating. *NST Online*. [Online]. Available: https://www.nst.com.my/news/nation/2020/03/575177/covid-19-movement-control-order-imposed-only-essential-sectors-operating

[6] Y. Palansamy. (May 2020). Higher Education Ministry: All university lectures to be online-only until end 2020, with a few exceptions. Malaymail.com. [Online]. Available: https://www.malaymail.com/news/malaysia/2020/05/27/higher-education-ministry-all-university-lectures-to-be-online-only-until-e/1869975

[7] UPSI to implement online teaching in line ministry's instruction. *theSundaily*. (2020). [Online]. Available: https://www.thesundaily.my/local/upsi-to-implement-online-teaching-in-line-ministry-s-instruction-DB2175401

[8] M. D. I. B. Baba and G. J. Pendek, "Keberkesanan pengajaran dan pembelajaran dan kaitannya terhadap prestasi akademik pelajar uthm," 2009.

[9] B. Smith and C. Brame, *Blended and Online Learning*, Vanderbilt University Centre for Teaching, 2014.

[10] A. Nguyen, "The impact of online learning activities on student learning outcome in blended learning course," *Journal of Information & Knowledge Management*, 2020.

[11] A. Rajuladevi. (2018). A machine learning approach to predict first-year student retention rates at University Of Nevada, Las Vegas. *Digital Scholarship@UNLV*. [Online]. Available: https://digitalscholarship.unlv.edu/thesesdissertations/3315/

[12] V. Palaniappan. (2020). Predicting solar irradiance with SVM regression. *SSRN Electronic Journal*. [Online]. Available: https://www.academia.edu/38579134/Predicting_Solar_Irradiance_with_SVM_Regression

[13] A. Kesumawati and D. T. Utari, "Predicting patterns of student graduation rates using Na ïve bayes classifier and support vector machine," *AIP Conference Proceedings 2021*.

[14] J.-H. Fu, J.-H. Chang, Y.-M. Huang, and H.-C. Chao, "A support vector regression-based prediction of students' school performance," in *Proc. 2012 International Symposium on Computer, Consumer and Control*, Jun. 2012, doi: 10.1109/is3c.2012.31.

[15] E. Y. Obsie and S. A. Adem, "Prediction of student academic performance using neural network, linear regression and support vector regression: A case study," *International Journal of Computer Applications*, vol. 180, no. 40, May 2018.

[16] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, 1992.

[17] Y. Zhang and L. Wu, "Classification of fruits using computer vision and a multiclass support vector machine," *Sensors (Basel, Switzerland)*, 2012.

Nor Ain Maisarah Samsudin was born in Terengganu, Malaysia. She is a student of master's degree in applied statistics at Sultan Idris Education University (UPSI). In her studies, she focused on visualization, modelling and prediction of soil microbial community using machine learning approach coupled with multivariate analysis. Her studies also involved with the area of dimension reduction method to reduce dimensional data by based on techniques in Data Mining. Her research had been published in Scopus indexed journal and her work had been preented in local conferences.

Shazlyn Milleana was born in Johor Bahru, Malaysia, in 1988. She is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Sultan Idris Education University (UPSI). She graduated with a bachelor's science degree in industrial mathematics from Universiti Teknologi Malaysia, in 2010. Upon graduation, she commenced her career as an Executive in a banking institution. In the following year, she was offered to further her studies as a fast-track PhD student at the same university. During her PhD journey, she developed an interest in multivariate analysis, focusing on discovering patterns that deal with big data. Her research investigates the area of dimension reduction methods with a focus on climate informatics involving the analysis of big climate-related datasets via the n techniques in Data Mining. her research had been published in Scopus indexed journal and presented at numerous local and international conferences. Her PhD thesis was completed by end of 2016 and therefore she received a doctorate in 2017.

Nurul Ainina Filza Sulaiman was born in Selangor, Malaysia. She is a student master of applied statistics in Sultan Idris (Education University UPSI). In her studies, she focused on statistical downscaling by machine learning in classification and regression models in the East Coast Peninsula of Malaysia. Her studies also involved with the area of dimension reduction method to reduce dimensional data by based on techniques in data mining. She published her research in Scopus indexed journal and presented her work at local conferences.

Shuhaida Ismail is a lecturer at the Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia (UTHM). Her first degree is in computer sciences major in UTM. She also completed her master's degree and PhD from the same university. During her studies, she acquired an interest in Machine Learning research areas, specifically in predictive modelling, classification, and clustering. Her current research areas are hydrological modelling, big data analytics, and deep learning.

Nur Syarafina Mohamed is a senior lecturer at Universiti Teknologi Malaysia specialising in the Optimization area. She graduated with a bachelor's science degree in industrial mathematics from Universiti Teknologi Malaysia, in 2010. Upon graduation, her career as a lecturer in Universiti Teknologi Mara commenced from 2010 to 2016 on a contract basis. In 2013, she began her PhD journey at Universiti Sultan Zainal Abidin in Kuala Terengganu Malaysia. Her research interest is focused on Optimization where she modified the Conjugate Gradient Method by implementing new adjustments to the coefficient employed in the algorithm. The parameter introduced is compared among the best parameters that were introduced previously. She published her research in Scopus indexed journal and presented her work at various local and international conferences. She completed her PhD thesis in April 2017 and was conferred a doctorate in December 2017.

Nor Hafizah Md Husin was born in Kota Bharu, Kelantan. She is currently a senior lecturer at the Mathematics Department, Faculty of Science and Mathematics, Sultan Idris Education University, Tanjong Malim, Perak. She obtained her first degree in Computational Mathematics from Universiti Malaysia Terengganu (UMT). She also obtained a Master's degree and Ph.D from the same university. She holds a Ph.D. in mathematical sciences specialising in graph theory. She published her research in Scopus indexed journal and presented her work at various local and international conferences.