A Prediction Model for Blood Donation Using Multiple Logistic Regression

*Wan Hanieza W. Mohamad Hanapi[1], Haslina Md Sarkan[2], Nilam Nur Amir Sjarif[3], Yazriwati Yahya[4], Suriayati Chuprat[5]

Razak Faculty of Technology and Informatics

Universiti Teknologi Malaysia
hanieza@dijohor.com.my[1], haslinams@utm.my[2], nilamnur@utm.my[3], yazriwati.kl@utm.my[4],suriayati.kl@utm.my[5]

Article history

Received:
9 Sept 2019

Received in revised form:
31 Oct 2019

Accepted:
20 Dec 2019

Published online:
30 Dec 2019

*Corresponding author
hanieza@dijohor.com.my

Abstract

Minimal participation in blood donation is a concern in this country despite many blood donation programs organized. This study attempts to find out the factors influencing the intention level of people to donate blood. The main objective of this paper is to identify the association between the willingness of donating blood with the number of months since the last donation, number of donation, total volume donated and the number of months since the first donation. Secondary data retrieved from UCI Machine Learning Repository were used. Based on the Logistic Regression Model, there are only three factors affecting the willingness to donate blood which are month since last donation, frequency of donation and total volume donated. There are four ways to validate the model which are Hosmer and Lemeshow Test, Coefficient of Determination, Classification Table and area under Receiver Operating Characteristics (ROC) value. The validation tests showed that the final model has a good performance.

Keywords: *Prediction model, machine learning, Logistic Regression Model.*

1. Introduction

Red blood cells are important for blood transfusion that is usually required in cases of accidents, operations as well as for thalassemia patients who are need almost at least one transfusion per month. To meet with the local demand, it is the responsibility of the blood centre to manage the blood stock wisely. However, it is not an easy task since overstocking and under stocking of blood might occur. One of the factors that lead to under stocking is the lack of donor (Lestari et al., 2019). Many people are still not aware of the importance of donating blood. In fact, many people believe in the myth of blood donation such as increase in weight, weakening of the body and easily infected by diseases (Gebresilase, Fite & Abeya, 2017).

However, overstocking of blood can happen when there is expired blood stock. This situation is also not good for the blood centre because the expired blood needs to be disposed, where there is high cost of getting rid of the blood and the task is not easy. Therefore, it is very important for the management of the blood donation to know the landscape of blood demand and the forecast value. In this paper, we propose the logistic regression model to predict the number of possible blood donation in order to determine surplus or shortage of blood.

** Corresponding author. hanieza@dijohor.com.my*

It is then essential to identify the association between the willingness to donate blood with the number of months since the last donation, number of donations, total volume donated and number of months since the first donation.

The aim of this paper is to predict the potential of blood donors based on his past donation behaviour. This is important in order to help the blood centre to plan a deliberate strategy such as organizing more interesting blood donation campaigns to increase awareness of the importance of donating blood. Besides that, this paper is also trying to determine how strong the significant factors that affect the willingness of donating blood.

The research questions derived from the main objective are:

1. What is the relationship between the willingness of donating blood with the number of month since last donation, number of donation, total volume donated and number of month since first donation?
2. How are we going to apply Multiple Logistics Regression in determining significant factor that affects the willingness to donate blood?

In order to answer the above questions, we divide this paper into several sections. The first section introduced the existing blood donation problems and we believe that predicting potential blood donators is the issue that we should focus on. The following section will present the literature work done to further understand the chosen issue. We will discuss about the methodology used in section 3. Section 4 will discuss about the work done and we will conclude the entire work in Section 5.

2. Literature Review on blood donation

As to date, there is no safe substitute of blood despite the many researches done, as they come with significant adverse effects (Health, 2012). The sources of blood are very limited since blood only comes from the donors. Therefore, blood donation is very crucial as it can save many lives.

2.1. Identifying Blood shortage and surplus

Most blood donation centres face blood shortage. Therefore, it is important to identify the factors that can lead to the shortage of blood. Firstly, blood shortage happens because of poor management of donated blood and this may lead to blood wastage (Kumari & Wijayanayake, 2016). In 2010, World Health Organization (WHO) reported that 87.5% of developing countries are not able to collect blood stock more that than half of the actual demand. It is very critical that even the lack of blood stock is only at 13.5%, it can harm a life.

Recent study by Cheah & Tang (2017) revealed that only 2.2% of the total populations are willing to donate their blood in Malaysia. Hence, he studied the factors that cause low of blood donation rate by studying sociodemographic and lifestyle on various donor status. The study has reported percentage of blood

* *Corresponding author* *haslinams@utm.my*

donation by male is higher than female. This might be because women cannot donate blood during pregnancy stage as well as breastfeeding periods.

Blood wastage on the other hand, is referring to any blood component or blood product that is discarded rather than administered to a patient. There are several reasons that lead to blood wastage. According to Sharma et al. (2014), the number of discarded bags caused by transfusion transmissible diseases (TTI) can be reduced by strictly filtering the donators. Staff at the blood centre should carefully look at the criteria of donator by asking several important questions to them before they can donate their blood.

Kurup et al. (2016) asserted that the other factor that led to blood wastage was poor blood storage processes such as broken bag and clotted blood. They conducted a study on blood product usage and wastage at Georgetown Public Hospital Cooperation (GPHC), Guyana and found that the percentage of blood wastage there from 2012 to 2014 was 25%. The study revealed poor quality in managing blood stock as the main reason of the wastage. The successfulness of blood quality management depends on several factors which are careful collection, accurate testing, processing and labelling as well as efficient storage and distribution process (WHO,2010).

Far et al. (2013) also found that 77.9% of wasted blood in Iranian hospitals was caused by expiry date. Packed red cell has the highest wastage level which is 59.4% while the least contributor to wastage is in the form of cryoprecipitate (2.4%). The rest are plasma (22%) and platelet (16%). In the meantime, it was found that 58.3% of the packs were wasted in teaching hospital. The least wastage is from military hospitals (1.9%). The discarded unit in social welfare and charity hospitals are 15.8% and 3.2% respectively.

A study by Morish et al. (2012) found seropositive as the one of blood wastage factor. However, it is the least contributor (5.8%) compared to expired blood (27.4%), leakage (25.7%), lipemic (11.4%) and suboptimal collection (16.1%). The high percentage of discarded blood will affect the cost of waste disposal. Hence, it was suggested that proper interviews should be made before the donor can proceed with the blood donation process. Overall, it can be concluded that expired unit of blood is the most popular reason of the wastage.

2.2. Blood donation awareness

The awareness of blood donation in Malaysia is still low. It was reported that the number of blood donors was only approximately 2.2% to 2.5% of the entire population in 2017. This number was very far too low compared to other developed country which was at 5%. Lim et al. (2018) revealed that one of the factors of blood shortage in Malaysia was because the growth in donor numbers was contributed by repeat donors instead of new donors. Besides that, the National Blood Centre faced the dropping number of blood donation during festival seasons and school holidays.

2.3. Blood forecasting challenges

One of the challengers in forecasting blood is that blood is a perishable item which has limited shelf life (Fortsch & Khapalova, 2016). Besides that, blood is also considered as perishable item because of the tendency of it being mismatched between the blood donor who acts as supplier and blood recipients who are demanding the blood.

2.4. Proposed solution based on dataset and literature

Four factors are proposed in this paper which are number of months since the last donation, the number of donations, the total volume donated and the number of months since the last donation. These factors are proposed based on the dataset used for this study. According to Stewart (2019), statistics and machine learning are distinguished by their different purposes. Statistical models are designed for inference about the relationships between variables. Meanwhile, machine learning models are designed to make the most accurate predictions possible. In this work, we are interested to analyse the factors that lead to the willingness of donating blood and this led to the choice of using Multiple Logistics Regression.

In the meantime, we believe that the four factors are relatable with the profile of the donors which are new donors or repeated donors based on the history of data. Lim et al. (2018) also found that the percentage of donation by repeat donors was rapidly growing between 2008 to 2014 by 37.1% in Malaysia. This is much higher compared to blood donation by new donors, 13.3%. Experienced donors tend to re-donated their blood as their past donation frequency increased (Godin et al., 2007).

3. Methodology

Logistic regression is a regression model where the dependent variable is categorical or binary dependent variable and the independent variables are continuous. The scope of this paper exclude the model’s performance evaluation.

3.1. Data Collection

The data was retrieved from UCI Machine Learning Repository. This dataset gives information about blood donation by staff and students from university in Hsin-Chu City. There are 748 instances and five variables including a response variable. Independent variables consist of the Number of Month since Last Donation, Number of Donation, Total Volume Donated and Number of Month since First Donation while response variable is the Willingness of Donating Blood. It is cheaper and faster using secondary data. We save time, effort and money. Secondary data can help to creating new perception from previous analyses because reanalysing data can result in unanticipated new findings.

3.2. Pre-processing of Data

In this study, it is essential to make sure there is no missing value. This is because it will affect the whole prediction process. We had to decide for the most suitable method to fill up the missing values. There are four ways to handle the missing values. Firstly, the author can simply delete the missing value. Next, missing value can be replaced by average value. Besides that, it can be integrated by using isolation. Lastly, the author can take either the value before or after the missing value. Since there is no missing value in the dataset, we can proceed to the analysis process.

3.3. Descriptive Analysis

Descriptive statistics is a set of brief descriptive coefficients that recapitulates a given data set. The whole population can be represented by this set, or only a sample. In order to describe a data set, we use measures of central tendency and measures of variability or dispersion. Mean, median and mode describes the measure of central tendency while standard deviation or variance, the minimum and maximum of the variables, kurtosis and skewness for measure of variability and dispersion.

3.4. Simple Logistic Regression

Simple logistic regression denotes estimation of the association between dependent and independent variable. This univariable analysis is used to screen the important independent variables.

3.5. Multiple Logistic Regression

An estimation of relationship between dichotomous dependent variable and two or more independent variables make up for multiple logistic regressions. We apply the forward selection method here. We add the independent variables to the model one at a time until none of the remaining variables are significant to be added into the model.

3.6. Checking Multi-collinearity

Multi-collinearity is a where two or more independent variables are highly correlated to each other. In general, when two or more independent variables are related to each other, it provides redundant information about the dependent variable. Variance Inflation Factor or VIF measures how much the variance of the estimated coefficient is inflated as compared to when the independent variables are not linearly correlated. The multi-collinearity exists when the VIF is greater than 10 or tolerance is less than 0.2 or 0.1.

4. Results & Discussions

In this section we discuss the results and findings of this work order to achieve the main objective of the paper.

4.1. Descriptive Analysis

In this paper we describe the quantitative variables by using minimum, maximum, median and mean value. Besides that, we also use boxplot to describe the factors.

4.1.1. The Willingness towards Donating Blood

Figure 1 shows the willingness of previous blood donators to donate their blood in March 2007. Only 178 (23.8%) of the donators want to donate their blood while the rest 570 (76.2%)choose not to at that particular time.

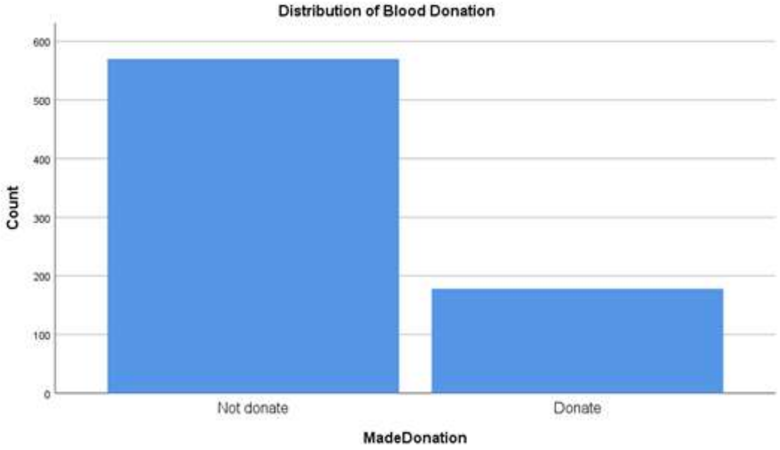

Figure 1. The Willingness to Donate Blood

4.1.2. Months since the last Donation

Figure 2 shows the boxplot of months since the last donation. The maximum number of total months since the last donation is 74. This indicates that the longest time difference between the previous and current donation is six years. Mean equal to 9.51 indicates that most of donators have been donated blood for the past 9 months. They are translated in Table 1.

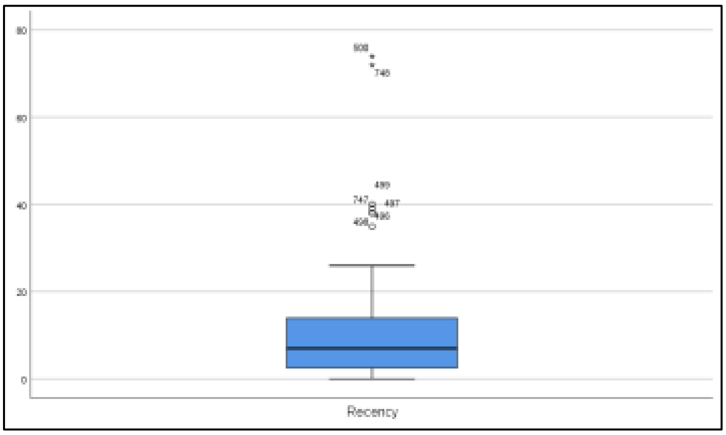

Figure 2. Boxplot of Months since Last Donation

Table 1. Min, Median, Mean and Max of Recency (Months)

Minimum	Median	Mean	Maximum
0	7.00	9.51	74

4.1.3. Frequency of Donating Blood

Figure 3 shows the boxplot for frequency of donating blood. The least frequency is one indicating that there are people that only have been donated blood once. In the meantime, maximum value is 50. This is mean that the highest number of donating blood by a person is 50 times as shown in Table 2. Averagely, people from the university tend to donate 5 times.

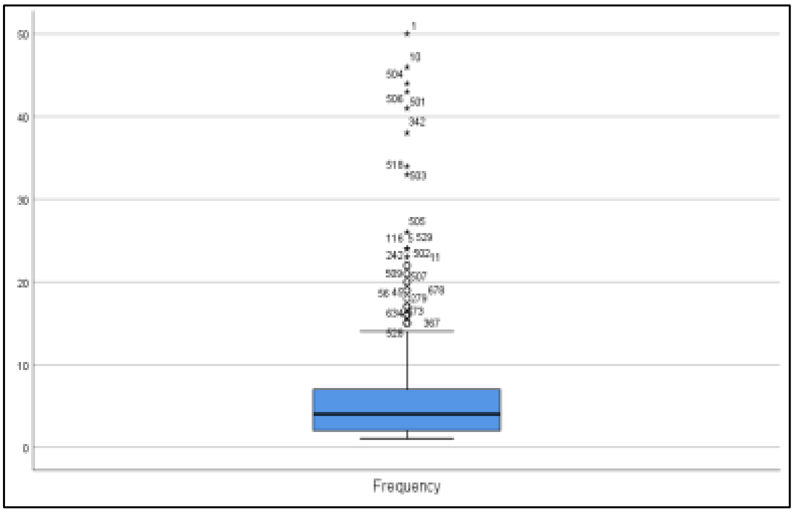

Figure 3. Boxplot of Frequency of Donating Blood

Table 2. Min, Median, Mean and Max of Frequency of Donating Blood

Minimum	Median	Mean	Maximum
1	4.00	5.51	50

4.1.4. Total value of donated Blood

Figure 4 shows the boxplot of total volume donated blood. People from the university have been donating their blood 1378 cc averagely. The least volume of blood donated is 250 cc while the highest volume is 12500 cc as seen in Table 3.

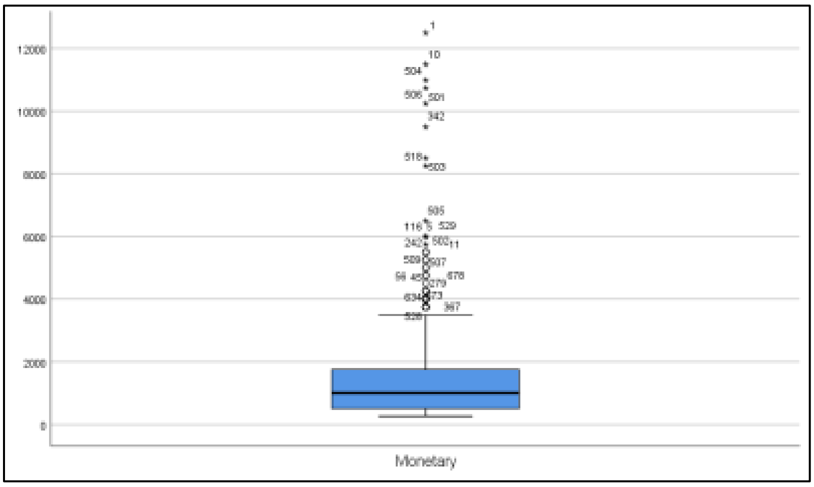

Figure 4. Boxplot of Total Volume of Donated Blood

Table 3. Min, Median, Mean and Max of Volume of Donated Blood

Minimum	Median	Mean	Maximum
250	1000	1378.68	125000

4.1.5. Months since First Donation

In average, the number of months since the first donation is 34.28 months. This indicates that most of the donors are new donors. The minimum number of first donation made is 2 indicating that there are new donors in the past two months. In the meanwhile, maximum value is 98 indicating that the most experienced donor has started donated his blood approximately eight years ago.

4.2. Logistic Regression

4.2.1. Multiple Logistic Regressions by Using Model Forward Selection Method

Based on the Multiple Logistic Regression in the following table, after one step using model forward selection method the three factors which are Total Months since the Last Donation, Frequency of Donating Blood and Total Volume of Donated Blood were found to be significant in the final model.

Table 4. Multiple Logistic Regression of Factors that Affect The Willingness To Donate Blood

Factors	B	Wald	P-value	Exp(B)	95% C.I.for Exp(B)	
					Lower	Upper
Total months since last donation	-0.099	32.409	0.000	0.906	0.876	0.937
Frequency of donating blood	0.135	27.813	0.000	1.145	1.089	1.204
Total volume of donated blood	0.023	14.491	0.000	0.977	0.966	0.989

Total months since last donation	-0.450	6.313	0.013	0.638		

4.2.2. Multi-Collinearity Test

Multi-collinearity exists if tolerance value is smaller than 0.1 and VIF value is more than 10. The presence of the multi-collinearity can be detected by the F-value in ANOVA is significant but insignificant of t-value of independent variables. The variance tends to be large and lead to loss of statistical. Using SPSS, the multi-collinearity can be assessed by Variance Inflation Factor (VIF) and Tolerance. Hence, it can be concluded that there is no multi-collinearity between independents variables since tolerance values are higher than 0.1 and VIF values of each factors are smaller than 10. We show the results in the following table:

Table 5. Multi-collinearity Test

Factors	Collinearity	
	Tolerance	VIF
Total months since last donation	0.838	1.193
Frequency of donating blood	0.514	1.947
Total volume of donation	0.518	1.932

For Simple Logistic Regression, three factors which are total months since last donation, frequency of donating blood and total volume of donated blood are significant towards the willingness of donating blood.

In Multiple Logistic Regression, Forward Selection Likelihood Ratio Method has been used by the author to select the best variables for this model. The result has shown that three variables or factors were selected due to the significant values that lower than 0.05.

Next, we tested the existence of multi-collinearity in order to identify either there is strong correlation between independent variables. This is important process since the performance of logistic regression model can be affected if there is multi-collinearity among the independent variables. Both Tolerance and VIF values must passed the requirement. Since the Tolerance value is higher than 0.1 and VIF value is less than 10, it can be concluded that there is no multi-collinearity between the factors.

4.2.3. Final Model

The final model for the multiple regression model is

$$P(y = 1) = \frac{1}{1 + e^{-z}}$$

Where;

z = - 0.450 – 0.099 Total months since last donation + 0.135 Frequency of donating blood + 0.023 Total volume of donated blood

P (probability = 1). (0 = not donate, 1 = donate)

The interpretation of the final model is as below:

1. If the total month since last donation is decrease by one year, it is 0.906 more likely for the donors to donate blood in March 2007.
2. If the frequency of donating blood is increased by one time, it is 1.145 more likely for the donors to donate blood in March 2007.
3. If the total volume of donated blood is increase by 1 cc, it is 0.977 more likely for donors to donate blood in March 2007.

5. Conclusions

The data was extracted from UCI Machine Learning Repository. We are interested to determine the factors that affect the willingness of donating blood among staff and students from Chung-Hua University. Based on the objectives, we found that the most suitable methodology to analyze the information is Logistic Regression method. The test of relationship between the independents and response variable was carried out to run the Logistic Regression Method. It is important to identify the odd ratio (OR) in Logistic Regression. This is because the value of OR evaluate the relationship between factors and the response. OR values display the odds that an outcome will occur given particular factors.

Only three variables or factors are contributing to the willingness of donating blood which are total months since last donation, frequency of donating blood and total volume of donated blood. Meanwhile, another factor which is months since first donation is not significant to the willingness of donating blood. Hence only three factors can be used to predict blood donation. In the future we will need to evaluate in detail the performance of the model.

6. References

Cheah, Y.K., and Tang, C.F. (2017) Factors Influencing the use of Preventive Medical Care in Malaysia: Evidence from National Health and Morbidity Survey Data. *Asian Economic Journal*, 31: 119– 137.

Far, R. M., Rad, F. S., Abdolazimi, Z., & Kohan, M. M. D. (2013). Determination of Rate and Causes of Wastage of Blood and Blood Products in Iranian Hospitals, 161–167.

Fortsch, S. M., & Khapalova, E. A. (2016). Operations Research for Health Care Reducing uncertainty in demand for blood. *Operations Research for Health Care*, *9*, 16–28.

Godin, G., Bélanger-gravel, A., Eccles, M., & Grimshaw, J. (2008). Healthcare professionals ' intentions and behaviours : A systematic review of studies based on social cognitive theories, *12*, 1–12.

Kurup, R., Anderson, A., Boston, C., Burns, L., George, M., & Frank, M. (2016). RESEARCH ARTICLE A study on blood product usage and wastage at the public hospital , Guyana. *BMC Research Notes*, 1–6.

Lim, M. L., S. H. Thock, A. K. G. Tan and S. L. Gwee. 2018. Determinants of blood donation status in Malaysia: Profiling the non-donors, occasional donors and regular donors,2018

Morish, M., Ayob, Y., Naim, N., Salman, H., Muhamad, N.A. & Mohd Yusoff, N. (2012). Quality indicators for discarding blood in the National Blood Center, Kuala Lumpur.*Asian J Transfus Sci. 2012 Jan-Jun*; 6(1): 19–23.

Gebresilase, H. W., Fite, R. O., & Abeya, S. G. (2017). Knowledge, attitude and practice of students towards blood donation in Arsi university and Adama science and technology university: A comparative cross sectional study. *BMC Hematology*, *17*(1), 1–10.

Health, N. (2012). *Encyclopedia of Exercise Medicine in Health and Disease*. *Encyclopedia of Exercise Medicine in Health and Disease*.

Kumari, D. M. S., & Wijayanayake, A. N. (2016). An efficient inventory model to reduce the wastage of blood in the national blood transfusion service. *2016 Manufacturing and Industrial Engineering Symposium: Innovative Applications for Industry, MIES 2016*, (October), 1–4.

Lestari, F., Ulfah, U., Aprianis, F. R., & Suherman, S. (2019). Inventory Management Information System in Blood

Transfusion Unit. *IEEE International Conference on Industrial Engineering and Engineering Management*, 2019-Decem, 268–272.

Authors

Wan Hanieza binti W. Mohamad Hanapi received her MSc. Business Intelligence and Analytics from Universiti Teknologi Malaysia. She is currently an analyst/statistician at Digital Johor Sdn.Bhd. Her research interest is in the area of social media analysis, regression analysis and big data analytics.

Haslina Md Sarkan is a senior lecturer at Razak Faculty of Technology & Informatics, Universiti Teknologi Malaysia. She did her tertiary studies in France where she received her Master’s degree in Electronics, Electrotechnique & Automatic from Université de Montpellier II and is currently doing her PhD in Software Effort Estimation. Her research interest started with image processing and biometrics applications before expanding into software engineering fields and her current research projects include software project management and disaster risk management using machine learning. She is MSTB certified and an IEEE member.

Nilam Nur Amir Sjarif is a senior lecturer at Razak Faculty of Technology and Informatics, under the department of Advanced Infomatics, Universiti Teknologi Malaysia, Kuala Lumpur. She received her PhD in Computer Science from Universiti Teknologi Malaysia in 2015, in the area of Human Action Recognition with the Geometrical Feature Representation for Video Surveillance. She is active in the research field Image and Video Processing, Pattern Recognition, Machine Learning, Deep Learning, Watermarking and Big Data Analytics.

Yazriwati Yahya is Senior Lecturer at Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia. She received BSc and MSc degrees in Computer Science (Software Engineering) and currently a PhD candidate from Universiti Teknologi Malaysia. Her research interest includes Software Engineering, Social Networking Sites Adoption and Psychology of Technology Adoption. She currently active in research and development projects in the area of Big Data Analytics and Information Systems. She is a member of the Association of Information System and IEEE Computer Society.

Suriayati Chuprat is Associate Professor at Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia. She received BSc and MSc degrees in Computer Science (Software Engineering) and PhD in Mathematics from Universiti Teknologi Malaysia. In part of her PhD research, she was attached to the University of North Carolina, USA. She did a postdoctoral program at the University of York, UK. Her research interests include Software Engineering, Algorithms and Scheduling Theories, Real-time Systems and Parallel Computing. She currently active in research and development projects in the area of Big Data Analytics and Cyber Security. She is a member of the ACM Professional and IEEE Computer.