Ensemble-Learning Techniques for Predicting Student Performance on Video-Based Learning

Chin-Wei Teoh, Sin-Ban Ho, Khairi Shazwan Dollmat, and Chuie-Hong Tan

Abstract—**The transformation of education norms from face-to-face teaching era to the Massive Open Online Courses (MOOCs) era has promoted the rise of the big data era in educational data. This situation has created an opportunity for an educator to utilize the available data from MOOCs to facilitate student learning and performance. Therefore, this research study aims to introduce three types of ensemble learning methods, which are stacking, boosting, and bagging, to predict student performance. These techniques combine the advantage of feature selection method and Synthetic Minority Oversampling Technique (SMOTE) algorithm as a method to balance the number of output features to build the ensemble learning model. As a result, the proposed AdaBoost type ensemble classifier has shown the highest prediction accuracy of more than 90% and Area Under the Curve (AUC) of approximately 0.90. Results by AdaBoost classifier have outperformed other ensemble classifiers, stacking and bagging as well as base classifiers.**

Index Terms—**AdaBoosting, ensemble learning, educational data mining, SMOTE algorithm.**

I. Introduction

Previously, having educators physically present to teach has been regarded as the main way to transmit knowledge to students [1]. However, a new educational norm has been created due to the emerge of the educational technology domain where many students have accessing e-learning web platforms such as Edpuzzle, Coursera, and Udemy for learning new knowledge [2].

According to the MarketDigits forecast report, the MOOCs market is forecasted to have a CAGR (Compound annual growth rate) of 32.8% from 2021 to 2026 [3]. The rise of MOOCs data has created an opportunity for researchers in the education data mining (EDM) to facilitate student learning performance. In this case, many researchers have applied the knowledge of artificial intelligence to develop a predictive model on MOOCs data to predict student performance [4]. It is beneficial for educators to analyse student-learning trends and collect some insight to help educators make better decisions when structuring teaching methods.

In the MOOCs system, video was a primary source of MOOCs course design where every learning course will be uploaded by at least one video form learning material [5]. As mentioned by inventor of Cone of learning theory, Edgar Dale, people generally remember 10% of what they READ, 20% of what they HEAR, 30% of what they SEE, 70% of what they SAY and WRITE, and 90% of what they DO [6]. Moreover, videos cover both the HEAR and SEE part which able to let learner to remember at least 50% of contents delivered.

A report from YouTube Marketing has claimed that in January and February 2020 where the time of covid-19 pandemic, just over 300 educational videos were uploaded to YouTube with remote teaching or distance learning in the title and in March 2020 alone, that number was increased over 23,000 [7]. Moreover, growing of video-based learning has create the opportunity of learning analytic on educational video.

However, educators find it difficult to identify the learners at risk of poor performance in a timely manner due to the increasing size of MOOCs data. For this reason, educators need a quick and accurate automatic predictive system that to scrutinize the MOOCs data and deliver a prediction on individual student performance after went through MOOCs system. Therefore, this research paper aims to develop a predictive model using MOOCs data to identify student performance on video-based learning.

II. Related Work

Educational research has indicated that watching all assigned educational videos on a single day tend to induce heavy study load on the students. Moreover, many students in this study likely to waited until the last day of assignment to watch it and tend to skip less important content in video to speed up the learning process [8]. Another study also has indicated that an increase in educational video views on a day before tests and assignments in video-based learning [9]. At the same time, an increase in video watching view before test and assignment has inspired that total time of the student spent on watching video must be measured.

Furthermore, a study which surveyed total of 357 papers in student performance has listed features were mainly impact to learner performance such as student engagement, demographics, and psychomotor skills [10]. Therefore, this research experimental data has focused to collect the demographic of users such as age, gender, and their previous academic performance, as well as learning style which are considered as factors to influence student performance in MOOCs.

Many researchers have applied the machine learning method to develop a predictive model for predicting student

Manuscript received January 5, 2022; revised March 23, 2022.

Chin-Wei Teoh, Sin-Ban Ho, and Khairi Shazwan Dollmat are with Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia (e-mail: chinwei2060@gmail.com, sinbanho@gmail.com, shazwan.dollmat@mmu.edu.my).

Chuie-Hong Tan is with Faculty of Management, Multimedia University, 63100 Cyberjaya, Malaysia (e-mail: chtan@mmu.edu.my).

doi: 10.18178/ijiet.2022.12.8.1679

performance using MOOCs data. However, in prior years ago, ensemble learning methods have been outperformed individual machine learning techniques on the task of predicting student performance in the education domain. The research conducted by Xu et al. has indicated that adapting the ensemble learning technique is able to improve the predictive model performance [11]. In addition, another study also indicates that the approach of the heterogeneous ensemble learning method has performed better in predicting student performance than sole machine learning method [12].

In a study that related to video learning analytic, the author has applied a several machine learning algorithms such as random forest, naive Bayes, support vector machine (SVM) and logistic regression to classify student performance and the random forest model performance achieve approximately 88% of accuracy [13]. Furthermore, another study has implemented an ensemble learning technique called AdaBoost classifier which successfully improves the accuracy of the predictive model to approximately 80% on predicting student performance [14].

Feature selection has been a key of research whether possible to improve the performance of a predictive model. A combination of ensemble learning models and chi-square test as feature selection method has been a key of improving the accuracy of the predictive model [15]. Another study by Ebrahimi-Khusfi et.al also shows that the addition of the feature selection method improves predictive model performance [16]. Furthermore, handling of imbalance data method such as SMOTE has been applied together with ensemble learning algorithms to enhance the performance of predictive model [17]. A combination of feature selection and SMOTE also has been applied for high imbalance datasets to enhance the performance of predictive model [18].

Therefore, most of the previous related work shows that the ensemble learning method outperforms other sole machine learning methods in predicting student performance. In addition, feature selection and handling of imbalance data method are key factors in improving predictive model performance. However, most of the previous related work had focused on clickstream data engagement rather than considering learner's engagement based on each section of video such as number of views on each section of video. Moreover, this research work would go into direction on collecting number of views on each video section.

III. Methodology

A total of 110 students have been assigned to watch an educational software design type of video content and then total of 22 features from video watch engagement and learning style have been collected into a complete dataset. Firstly, the original datasets first went through a feature selection method, Chi-Square test and then SMOTE algorithm to solve the imbalance data target in the original datasets. Next, grid search as a hyperparameter tuning method has been applied to search for optimal parameters to build an ensemble learning model. The datasets were split into 30% for test set and 70% for training set. Three different types of ensemble learning methods which are stacking, boosting, and bagging were built.

Finally, the performance of all ensemble learning methods and base classifiers has been evaluated based on the Receiver Operating Characteristic (ROC) curve and metrics such as accuracy and the Matthews Correlation Coefficient (MCC). The final prediction of the model was a binary classification where *1* indicates the student score full mark and *0* indicates the student fail to score full mark.

A. Overall Application Framework

As illustrated in Fig. 1, there are two categories of input features has been collected which were video engagement features and student 's learning style. In the MOOCs system or e-learning system, the number of views on each section of video has been collected. At the same time, total time spent by student to watch an educational video also has been collected.

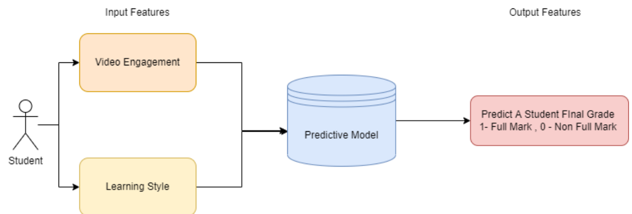

Fig. 1. The overall application framework.

The determination of learning style has been collected via online questionnaire. In this case, Felder Silverman (FS) learning style dimension has been applied to distinguish student leaning style. After that, all data has been used to build a predictive model and then predict a student final grade in the online course whether will get full mark. This predictive application was aim for small-scale video-based learning course which allow the educator to identify those learners who can score full mark at the end of the course.

B. Dataset Description

A total of 22 data attributes from original datasets. In the total of 22 features, 2 features are nominal data types while 20 features are integer data types. The nominal data types have provided non-numeric values such as a label. As shown in Table I, all 21 input features except Grade which as the output can be categorized into three different categories which are student demographic background, student video watch engagement and student learning style features.

TABLE I: Data Categories Descriptions

Data Categories	Data Features
Student demographic background	Gender, CGPA_class
Student video watch engagement	Time_1, Time_2, Time_3, Time_4, Time_5, Time_6, Time_7, Time_8, Time_9, Time_10, Grade_Q1, Grade_Q2, Grade_Q3, Rating
Student learning style	A/R Score, S/I Score, Vi/Vb Score, S/G Score

C. Base Classifiers

A supervised machine learning algorithm, the Support Vector Classifier (SVC), is used for classification. The SVC uses a technique called the kernel trick to transform the training data to identify an optimal boundary between the possible outputs. K-Nearest Neighbour (KNN) is another supervised machine learning algorithm for solving a classification problem. It determines an optimal factor *K* value can be determined through measuring training and the validation error rate in the experimental test. The *K*-value which has the lowest error rate would be selected as the optimal *K* value.

The prediction output from various decision trees is combined using an ensemble learning technique called the Extremely Randomized Trees (Extra Tree). It creates many unpruned decision trees from the training data, then takes the majority vote of those trees to create the final prediction [12].

D. *Ensemble Learning Methods*

Stacking is one of the ensemble machine learning algorithms that combine different types of machine learning base classifiers to produce more accurate predictions [12]. At first, the 10-fold cross-validation technique is applied to separate the training set and train all base classifiers. After that, the feature vector is used to predict the outcome prediction from each base classifier and generate a new set of training features to build the meta-classifier model. The test data will be applied for the meta-classifier model to perform the final prediction.

The Boosting ensemble learning method works to trains several decision tree models and then aggregates it to the ensemble model. The boosting mechanism will update the weight of the training set based on each ensemble model until reaching the phase of producing the last ensemble model is produced [14].

In this paper, adaptive boosting, or called AdaBoost, has been selected. At first, AdaBoost assigns weights to each training set after training a weak classifier model. Then it updates the weight of all mis-classifiers will be updated. In this case, the weight of each training set will be updated with the following equation (1):

$$H(x) = sign(\sum_{t=1}^{T} a_t h_t(x)) \tag{1}$$

$h_t(x)$ is the output of the weak classifier t for input, a_t is the weight assigned to classifier, a_t is calculated by 0.5 * ln ((1 — E)/E): it is based on the error rate E. The following equation (2) updates each training weight, where D_t is the weight at the previous level. In this case, the weights will be normalized by dividing the sum of all the weights, Z_t .

$$D_{t+1}(i) = \frac{D_t(i)\exp(-a_t y_i h_t(x_i))}{Z_t} \tag{2}$$

Bagging works by combining multiple base classifiers and then training them separately to produce a strong and accurate model [19]. Bagging ensemble learning first creates multiple bootstrap samples to act as independent datasets and then fits a weak classifier model for each of these training samples. The weight of the model will be aggregate into the final ensemble model.

There are two techniques to aggregate multiple models under classification problems: hard voting and soft voting. In the case of hard voting, the prediction output from each base classifier can be used as a vote, and the final bagging classifier model outputs the class that receives the highest number of votes. In soft voting, the final bagging classifier model outputs the model which produces the highest probability value.

IV. Results and Discussions

A. *Outcome of Data Preprocessing*

In the data preprocessing, the original dataset went through the process of checking of null value where there is no null value detected among all 110 entries with 22 features in the original datasets. Next, the two nominal features which are Gender and CGPA Class features have been converted into numerical data types as shown in Table II.

TABLE II: Outcome of Data Types Converted

Data Features	Categorical to Numerical
Gender	Female = 0, Male =1
CGPA Class	2.00 - 2.66 = 1, 2.67 - 3.32 = 2, 3.33 - 3.66 = 3, 3.67 - 4.00 = 4

B. *Outcome of Exploratory Data Analysis*

In this research, the number of watched on each video interval has been collected as shown in Table III. In this case, there are 10 video time intervals with each 31 minutes. Each video time interval consists of different video delivered contents. Maximum, minimum and average number of watched are collected based on each video time interval. From Table III, video interval between 2 min 4 sec until 2 min 35 sec has recorded the highest number of views among all video time interval. Each of video time interval has been ignored that mean skipped by student. In average, all video time interval has recorded at least one view.

TABLE III: Number of watched on Each Video Time intervals

	Video Time Interval	Video Delivered Content	Maximum Watched	Minimum Watched	Average Number of Watched
1	0:00 -0:31	Introduction of video theme	6	0	1.3
2	0:31 - 01:02	Introduction of Singleton	7	0	1.3
3	01:02 - 01:33	Introduction of Singleton	5	0	1.2
4	01:33 - 02:04	How to Implement Singleton	4	0	1.2
5	02:04 - 02:35	Explain of singleton class constructor	**9**	**0**	**1.6**
6	02:35 - 03 : 06	Explain of singleton class variable	4	0	1.1
7	03:06 : 03:37	Implementation of GetInstance class	5	0	1.2
8	03:37 - 04:08	Concepts Lazy Creation	5	0	1.2
9	04:08 - 04:39	explain of Lazy Creation implementation	5	0	1.1
10	04:39 - 05:11	Example of Singleton concepts in application and summary of video	6	0	1.3

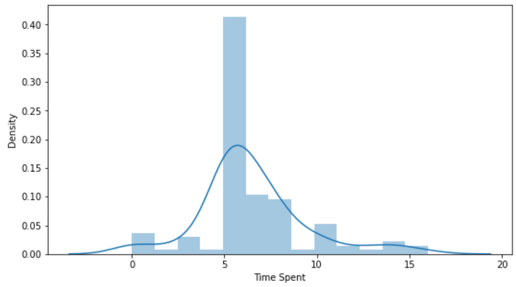

Fig. 2. Distribution on time spent variable on the video watched.

By observing Fig. 2 outcomes, most of the students are spent average 6 minutes to watch whole video. In this case, the actual length of the video assigned was 5 minutes 24 seconds. Meanwhile, the maximum length of video watched has achieved until 10 minutes, mean that perhaps student has watched some of part of video interval repeatedly.

Fig. 3 shows that the video time spent variable has been compared with student grade where 0 is non-full mark and 1 is score full mark. As a result, those students who spent more than 5 minute or repeat to watch the video has higher tendency to score full or excellent grade at the end compared to those who spent 5 minute and less to watch the video. This trend of analysis has shown that students have obtained sufficient information and knowledge to prepare on the quiz if they have spent longer time to watch the video.

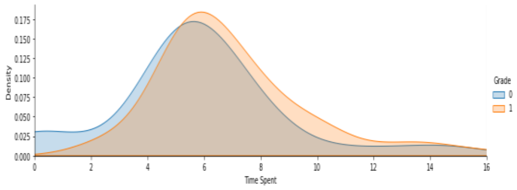

Fig. 3. Distribution on time spent variable on the student grade.

C. Outcome of Feature Selection

Fig. 4 shows that p-values calculated from Chi-Square test on each input features. In this case, the p-values compares independence between input features and output feature, *Grade*. Therefore, there are total of 11 input features has been selected where they found to have p-values less than 0.5 and significant dependent to the output feature as illustrated in the green zone in Fig. 2.

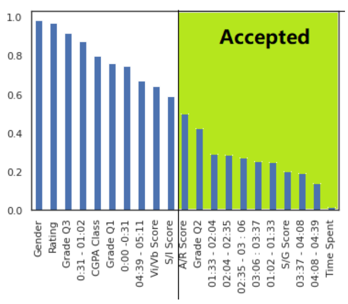

Fig. 4. P-values on each input features.

In the outcome of handling imbalance data target, the original datasets have total of 65 students who score full mark and 45 students who fail to score full mark. Therefore, SMOTE as an oversampling method has been applied to create balanced data target for both classes to contain 58 students.

D. Comparison of the Ensemble Learning Models Performance Results

Table IV shows that the AdaBoost ensemble classifier achieved the highest accuracy of approximately 90.9%, followed by Stacking (88.1%) and bagging (84.8%). In addition, the AdaBoost classifier model has achieved the highest MCC rate of 80.9%, showing itself to be the most effective predictive model for students who score under full and non-full mark class. The stacking classifier also achieves higher MCC rate (74.6%), making it the second effective model to classify the samples after the AdaBoost classifier model. As for the base classifiers, none of them were able to achieve higher accuracy than the stacking classifier. In this case, SVC and ET have the similar highest accuracy of 87.9%, followed by KNN which both achieve an accuracy rate of 86.5%.

TABLE IV: MODELS PERFORMANCE RESULTS

	Methods	Accuracy (%)	MCC (%)	AUC
Ensemble	Stacking	88.1	74.6	0.86
Learning	**AdaBoost**	**90.9**	**80.9**	**0.90**
Methods	Bagging	84.8	68.3	0.82
Base	SVC	87.9	74.5	0.86
Classifiers	KNN	86.5	73.4	0.85
	ET	87.9	74.5	0.86

*Note: Bold fonts to show the best achieved results.

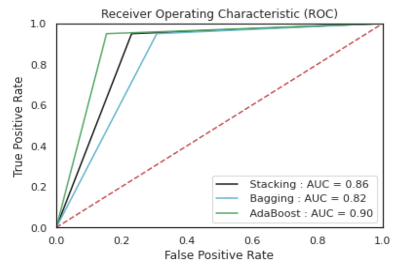

Fig. 5. ROC curve under different types of ensemble algorithms.

By observing Fig. 5 outcomes, the AdaBoost algorithm has a higher AUC rate of 0.90 than the stacking and bagging ensemble classifiers which are 0.86 and 0.82 respectively. Therefore, the AdaBoost ensemble learning classifier has achieved the highest AUC rate with the strongest ability to distinguish between the output classes.

V. CONCLUSION

This research study aims to develop and compare three ensemble-learning methods that integrate SVC, KNN, and ET base classifiers. The proposed methods also include the advantage of feature selection method, Chi-Square test and SMOTE technique to improve the performance of ensemble predictive models. The main findings are: Firstly, AdaBoost ensemble learning has the highest prediction accuracy of approximate 91% and AUC of approximately 0.90 compared to bagging and stacking ensemble algorithm. Secondly, the prediction performance of all three different types of ensemble learning methods outperforms all base classifiers. In future work, other effective type of ensemble learning methods and deep learning techniques is recommended to be explored and compared with current performance. Furthermore, genetic algorithm can be selected as another hyperparameter tuning option to optimize the ensemble learning models.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

CWT and SBH conducted the research and wrote the paper. KSD and CHT edited and improved the paper. All authors have approved the final version of the paper.

Acknowledgment

The authors appreciate the financial supports provided for this study by the Fundamental Research Grant Scheme (FRGS/1/2019/SS06/MMU/02/4).

References

[1] A. A. C. Hoyos and J. D. Velásquez, "Teaching analytics: Current challenges and future development," *IEEE Journal of Latin-American Learning Technologies (IEEE-RITA) IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 15, no. 1, pp. 1-9, Feb. 2020.

[2] B. Kurdi, M. Alshurideh, and S. A. Salloum, "Investigating a theoretical framework for e-learning technology acceptance," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 6484-6496, 2020.

[3] MarketDigits. (2021). *Massive Open Online Courses (MOOCs) Market Growing at a CAGR of 32.8% during 2021-2026*. [Online]. Available: https://www.globenewswire.com/en/news-release/2021/01/06/2154487/0/en/Massive-Open-Online-Courses-MOOCs-Market-Growing-at-a-CAGR-of-32-8-during-2021-2026-Growing-Popularity-Booming-Segments-Emerging-Trends-and-Investors-Seeking-Growth-Top-Players-Cou.html

[4] A. A. Mubarak, H. Cao, and S. A. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos," *Education and Information Technologies*, vol. 26, pp. 371-392, 2021, doi: https://doi.org/10.1007/s10639-020-10273-6

[5] D. J. Lemay and T. Doleck, "Predicting completion of massive open online course (MOOC) assignments from video viewing behavior," *Interactive Learning Environments*, pp. 1-12, 2020, doi: https://doi.org/10.1080/10494820.2020.1746673

[6] K. Masters, "Edgar Dale's pyramid of learning in medical education: Further expansion of the myth," *Medical Education*, vol. 54, no. 1, pp. 22-32, 2020, doi: https://doi.org/10.1111/medu.13813

[7] FutureSource. (2020). *Social Media and Microlearning Combine to Expand Access to Education Content*. [Online]. Available: https://www.futuresource-consulting.com/insights/social-media-and-microlearning-combine-to-expand-access-to-education-content/?locale=en

[8] K. L. Nielsen, "Students' video viewing habits during a flipped classroom course in engineering mathematics," *Research in Learning Technology*, vol. 28, pp. 1-12, 2020, doi: https://doi.org/10.25304/rlt.v28.2404

[9] B. Ahn and D. D. Bir, "Student interactions with online videos in a large hybrid mechanics of materials course," *Advances in Engineering Education*, vol.6, no. 3, pp. 1-24, 2018.

[10] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, and S. N. Liao, "Predicting academic performance: A systematic literature review," in *Proc. the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018)*, Larnaca, Cyprus, July 2018, pp. 175-199.

[11] J. Xu, Y. Han, D. Marcu, and M. Van Der Schaar, "Progressive prediction of student performance in college programs," in *Proc. the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, AAAI Press, San Francisco, California, USA, Feb. 2017, pp. 1604-1610.

[12] U. Pujianto, W. A. Prasetyo, and A. R. Taufani, "Students' academic performance prediction with k-nearest neighbor and C4. 5 on SMOTE-balanced data," in *Proc. the 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, Dec. 2020, pp 348-353, doi: https://doi.org/10.1109/ISRITI51436.2020.9315439

[13] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Applied Sciences*, vol. 10, no. 11, article 3894, June 2020, doi: https://doi.org/10.3390/app10113894

[14] M. Kumar, G. Mehta, N. Nayar, and A. Sharma, "EMT: Ensemble meta-based tree model for predicting student performance in academics," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, *1st International Conference on Computational Research and Data Analytics (ICCRDA 2020)*, Rajpura, India, October 2020.

[15] Q. R. S. Fitni and K. Ramli, "Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems," in *Proc. 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Bali, Indonesia, IEEE, July 2020, pp. 118-124, doi: https://doi.org/10.1109/IAICT50021.2020.9172014.

[16] Z. Ebrahimi-Khusfi, A. R. Nafarzadegan, and F. Dargahian, "Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques," *Ecological Indicators*, vol. 125, article 107499, June 2021, doi: https://doi.org/10.1016/j.ecolind.2021.107499

[17] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471-1483, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.358

[18] M. W. Huang, C. H. Chiu, C. F. Tsai, and W. C. Lin, "On combining feature selection and over-sampling techniques for breast cancer prediction," *Applied Sciences*, vol. 11, no. 14, article 6574, July 2021, doi: https://doi.org/10.3390/app11146574

[19] Z. Fang, Y. Wang, L. Peng, and H. Hong, "A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping," *International Journal of Geographical Information Science*, vol. 35, no. 2, pp 321-347, 2021, doi: https://doi.org/10.1080/13658816.2020.1808897

Chin-Wei Teoh received the B.S. degree in computer science from the Multimedia University, Malaysia, in 2020, and he is currently pursuing the M.Sc. degree in information technology at the same university.

Sin-Ban Ho received the B.Sc. and M.Sc. degrees in computer science from the University of Science Malaysia, in 1998 and 1999, and the Ph.D. degree in information technology from Multimedia University, Malaysia, in 2008. He is currently a senior lecturer in the faculty of computing and informatics at Multimedia University, Malaysia. He is a senior member of the IEEE and IEEE Computer Society.

Khairi Shazwan Dollmat received the B.S. degree in information technology in 2007, and the M.Sc. degree in information technology in 2015, both from the Multimedia University, Malaysia. He is currently a lecturer in the faculty of computing and informatics at Multimedia University, Malaysia.

Chuie-Hong Tan received the B.S. degree in mathematics in 1999, and the master degree in applied statistics in 2001, both from University of Malaya, Malaysia and Ph.D. degree in management from Multimedia University, Malaysia in 2009. She is currently a senior lecturer in the faculty of management at Multimedia University, Malaysia.