Available at https://jtsiskom.undip.ac.id (16 July 2019)
DOI: 10.14710/jtsiskom.7.2.2019.83-88
Jurnal Teknologi dan Sistem Komputer, 7(2), 2019, 83-88

Predictive Adaptive Test with Selective Weighted Bayesian Through Questions and Answers Patterns to Measure Student Competency Levels

Tekad Matulatan*), Martaleli Bettiza, Muhamad Radzi Rathomi, Nola Ritha, Nurul Hayaty

Department of Computer Science, Faculty of Engineering, Universitas Maritim Raja Ali Haji
Jl. Politeknik, Senggarang Campus, Tanjungpinang, Kepulauan Riau, Indonesia 29100

How to cite: T. Matulatan, M. Bettiza, M. R. Rathomi, N. Ritha, and N. Hayaty, "Predictive Adaptive Test with Selective Weighted Bayesian Through Questions and Answers Patterns to Measure Student Competency Levels," *Jurnal Teknologi dan Sistem Komputer*, vol. 7 no. 2, 2019. doi: 10.14710/jtsiskom.7.2.2019.83-88, [Online].

Abstract - Computer Assisted Testing (CAT) system in Indonesia has been commonly used but only to displaying random exam questions and unable to detect the maximum performance of the test participants. This research proposes a simple way with a good level of accuracy in identifying the maximum ability of test participants. By applying the Bayesian probabilistic in the selection of random questions with a weight of difficulties, the system can obtain optimal results from participants compared to sequential questions. The accuracy of the system measured on the choice of questions at the maximum level of the examinee alleged ability by the system, compared to the correct answer from participants gives an average accuracy of 75% compared to 33% sequentially. This technique allows tests to be carried out in a shorter time without repetition, which can affect the fatigue of the test participants in answering questions.

Keywords *- adaptive computer test; Bayesian probabilistic; selective weighted IRT; pattern behavior*

I. INTRODUCTION

Presently, attention to the CAT exam model that is better able to measure the level of understanding of test participants began to be developed, much like in [1] applying NLP in compiling patterns of questions given to examinees. A suggested model for the application of a virtual agent in the test of depression management through the CAT interactive system was tested on students in [2]. While in [3] applying automatic assessment to complex work in the CAT system and in [4] making comparisons between conventional test models and non-adaptive CAT test models.

Jeong [5] compared the scores of Korean students on computer-based and paper-based test versions. Wan and Henly [6] in their study of CAT found that figural response items are similar to multiple-choice items in providing information and efficiency. Meanwhile, response items made offer more information than multiple-choice items but tend to provide less information per minute. Beng et al. [7], using adaptive item selection, Huebner et al. [8] discussed the factors that influence the accuracy of the classification of computer test cognitive abilities, Buuren and Enggen [9] using Latent-Class-Based for the selection of test items, Brossman and Guille [10] comparing the multi-stage model with the linear test in the certification exam, Belov [11] and Belov [12] applying combinatorial optimization in determining the Assembly Test (TA) questions, Boonsathorn [13] merged the new format C-Test, S-Test, as part of the Computer-Based English Language Competency Test (CB TEC). The reliability coefficient, facility index, CB TEC discrimination index are carried out on the test data. Pearson correlation and regression analysis were also used. The results show that both forms of CB TEC have high reliability, validity related to high criteria, and high face validity.

Hakami et al. [14] discuss the determinants of the success of CAT implementation facilitated by Adaptive Structural Theory (AST) to assess technology acceptance by students. The conceptual model is proposed along with theoretical discussion; leaving the development and evaluation of a practical framework for the future.

Jamaludin et al. [15] present an increase in the valuation model using RFID (Radio Frequency Identification) technology that is implemented in a Computer Based Test (CAT). Malek et al. [16] provides CAT security considerations and proposes a Petri net-based framework for checking parallel models with Petri net separation procedures.

However, MST could have long iterative and also in a result, some student could have more questions than the others. In this paper, we proposed a technique to assess student competency using multiple-choice questions in which the questions would be randomly selected and based on the temporary test score of the student. It uses Bayesian to select a question that more likely the student can answer correctly. Each of the questions has a weight that also is corrected each time the student successfully answers or failed to give the correct answer. The student competency level can be estimated in a shorter time, without exhausting student's stamina in a long hour exam.

*) Correspondence author (Tekad Matulatan)
Email: tekad.matulatan@umrah.ac.id

Submitted: 23 October 2018; Revised: 11 February 2019; Accepted: 29 April 2019; Published: 30 April 2019

II. Research Methods

This research is a time series experimental, which was tested on Engineering Faculty students majoring in Informatics in the period of the academic year 2017-2018 even semester and partial in odd semester academic year 2018-2019 (Figure 1). The types of questions used in the CAT system developed in this study only focus on the types of multiple-choice questions with only 1 correct choice. Each question is made in 3 different difficulty levels, which are hard, standard, and easy levels with weights of 4.5, 3 and 1.5 respectively. Test participants will not be informed about the level of difficulty of the question, because it can be used by the examinee to find out the answers given are correct or not. The questions have been through a review phase by colleagues or in a team to ensure that they do not have a biased understanding nor the complicated narrative. The questions were arranged in the form of a study topics group so that the system can provide a deeper analysis of the competency of the examinees based on the ability of each topic.

The questions were recorded in the database along with choices and answers and the weight level of the questions. Validation in the database implemented to ensure the matching of questions and their answer choices and can be successfully displayed randomly. The algorithm implemented to select the questions and to record the feedback response (answer) of the examinee. The selection of questions was initially done sequentially. In the sequential process, the problem was displayed sequentially and loaded from the database. The experiment implementation used random question techniques where there was a better chance of predicting the maximum ability of the examinee.

The selection of the question will consider the weight of the question which can decrease or increase in weight depending on the pattern of the correct and wrong answers from the overall performance of the previous participants. The weight of each level of difficulty of the question consists of Hard (4.5), Standard (3), and Easy (1.5). This weight was reevaluated when the participant answers right or wrong using Eq. 1 where i is the i-th of the question, k the difficulty level of the question (hard, standard and easy), B_{ik} the number of correct answers to i-th question of level k, S_{ik} the number of wrong answers to i-th question of level k, N_{ik} the total number of answers to i-th question of level k, ω_k the standard weight of the question level k, and $weight_{ik}$ the weight value of the i-th question of level k.

$$Weight_{ik} = \omega_k - \frac{B_{ik}}{N_{ik}} + \frac{S_{ik}}{N_{ik}} \quad (1)$$

The weight of questions can increase or decrease at a maximum of 1 point. Thus, it is not possible to exchange questions with different weights when the problem has decreased in weight to lower than their weight below the level, or the problem has increased the weight to higher than their weight above the level.

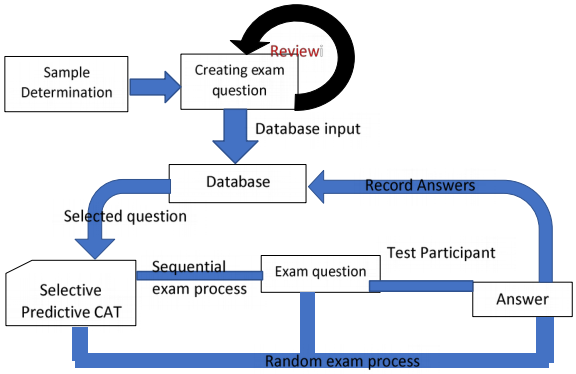

Figure 1. Experiment mechanism schema

Level	Weight
Hard	3.5
Standard	4

Level	Weight
Hard	4
Standard	3.5

Figure 2. Reevaluation of questions level

By reevaluating the weight of the questions, for example, when the type of problems is difficult, but all participants can answer this type, they experience a 1-point decrease to 3.5, while the questions in the same number with the standard type increase 1-point weight after none of the participants can answer right. Thus, there is an original level change of the problem, that is from difficult to a standard level and from standard to difficult level (Figure 2).

In order to minimize the chance of students passing by selecting the easy question, the minimum performance score was implemented. The assessment of the examinee is based on an assessment of performance with the minimal score threshold in Eq. 2, where N is the number of questions. For example with N = 10, then the pass threshold is 18.

$$P1 = (0.6*N*(Weight_{easy}1.5))+(0.3*N*(Weight_{standard}3))$$
$$P2 = 0.6*N*(Weight_{standard}3)$$
$$P3 = 0.4*N*(Weight_{hard}4.5) \quad (2)$$

With N = 10, the pass threshold of 18 is the minimum number to prevent cheating participants by deliberately choosing the wrong answer, so that the easiest question is given for the entire exam, then revise the answers to all the questions that have been being accepted. Although a participant has answered all 10 exam questions on easy levels, the performance score received is at maximum of 15. For reaching the score of 18, the minimum correct answer is at least 3 standard questions and 6 easy questions or at least 6 standard questions. This threshold limit can be changed but must consider the anticipation of the cheating by participants and the fairness of the completion value. The system proposed in this study does not apply a penalty system to wrong answers. To reduce the answers from the test

participants, the answer choices are made up to 6 variations to minimize the chance of guessing the correct answer.

The test participant's score performance is the achievement score of the results of the test participant's answers with the weight score of the questions. The performance score of the examinee is calculated using Eq. 3. Parameter i denotes the i-th of the question, k the difficulty level of the question (hard, standard and easy), B_{ik} the number of correct answers to i-th question of level k, N the total number of questions given, and $Weight_{ik}$ denotes the weight value of the i-th question of level.

$$score\ performance = \sum_{i}^{N} B_{ik} * Weight_{ik} \qquad (3)$$

The selection of questions by the system uses the adaptive model (Figure 3). The adaptive test modeling used in this study involves the pattern of answers from other participants before each question number and the pattern of answers from the target participants to determine the highest probability of answering the questions correctly. This is intended to be able to measure the maximum competency of the examinee. The probability of correctly answering the *i*-th question of level *k* is obtained in Eq. 4. Y denotes the class with the wrong or correct value, i the *i*-th of the question, k the difficulty level of the question (hard, standard and easy), and B_{ik} is the number of correct answers to *i*-th question of level *k*.

$$P(Y \vee Question_{ik}) = P(Y_{ik}).B_{ik} \sum_{1}^{i} P(Y_i) \qquad (4)$$

The test begins with the determination of the standard problem as in Figure 4(a). The initial question on the sequential model is problem number 1 of the sequence of questions, while in random questions it is determined at random results. The results of the participants' answers, although they can still be revised, determine the weight of the next question (Figure 4b). In this case, when the participant chooses the answer and moves to the next question, in the random model, the next question number is generated from the system but the difficulty level of the question that is displayed is in accordance with the conditions of the previous answer. Problems with the highest weight with the highest chance of being answered correctly are the choice of questions. As for the difference in the results of the answers from the test participants with the suggested questions, the level of accuracy of the system in predicting and classifying the competencies of the examinees.

The supporting tools used in this research are listed in Table 1. It applied web plaftorm using PHP language and MySql database. We uses Apache web server. Client can access the system using a browser supporting HTML 5.0.

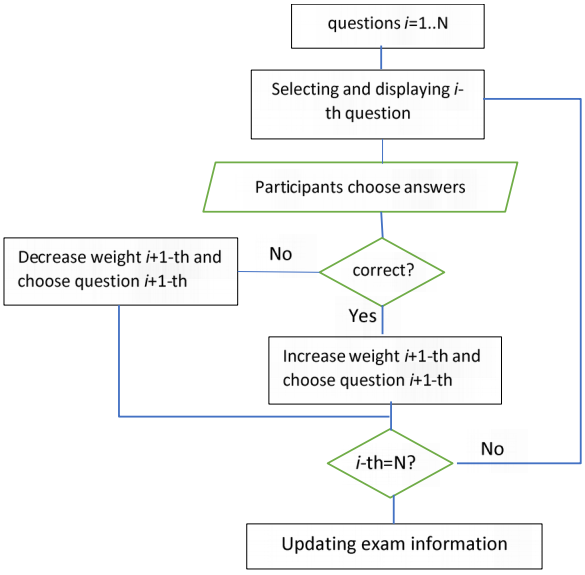

Figure 3. General process flow of adaptive test

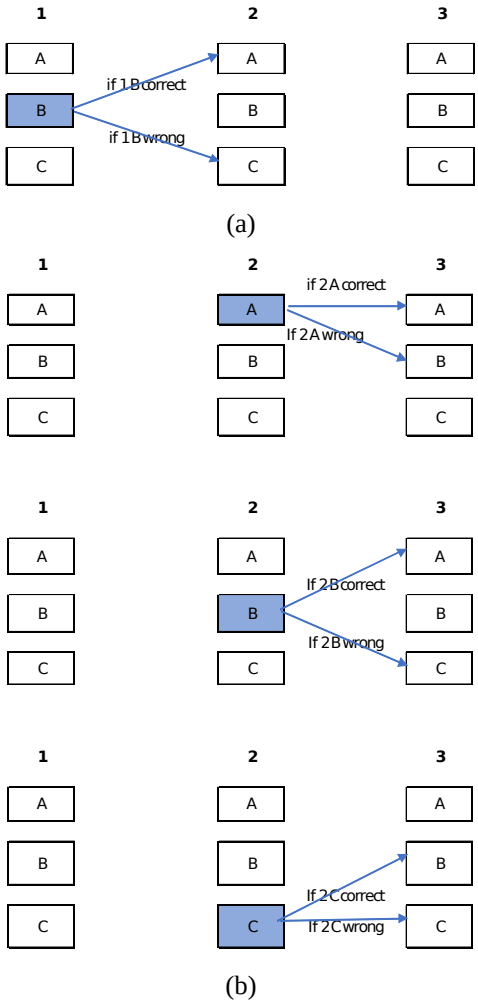

Figure 4. Test process: (a) Starting test with standard weight (b) Selection probabilistic after starting test

Table 1. Supporting tools

No	Description	Tools
1	Application platform	Web
2	Programming Language	PHP Alternative: JavaWeb
3	Application Server	Apache Alternative: Tomcat or Glasfish (for Java)
4	DBMS	MySql
5	System Architecture	Multitier with thin client
6	App Server pack	XAMPP
7	Database Editor	MySQL Workbench Community Edition
8	Client	A browser supporting HTML 5.0
9	License	GPL (General Public Licence)

III. Result And Discussion

The system, namely Selective Weight Bayesian (SWB), implemented in the subject sample population of Informatics Engineering majors, either sequentially or randomly with 10 questions. Figure 5 shows the experimental results on the sample by showing the accumulated number of correct answers given by the examinee at the difficulty level of each number. As expected, the problem with difficulty level easily dominates the number of correct answers, except for certain numbers as in question number 6, where the problem is more difficult than the standard and easy. This level of the correct answer only shows the tendency of the questions that appear and are answered correctly and does not describe the level of difficulty of the problem. This is due to the possibility of the problem rarely being chosen to display so that it has a low level and not because it is difficult to answer correctly.

The probability of answering correctly in each question with the level of difficulty has a relation to the level of correct answers (Figure 6). This probability is obtained from the comparison of the number of correct answers to the number of occurrences of the question. This level of probability provides a general description of the level of difficulty of the question. As in number 6, the probability of answering a difficult question is higher than the easy one, while in problem number 3, the opportunity to answer the standard problem correctly is higher than the difficult one. A case occurs in question number 5 of standard level which shows the number 0. This requires special handling where the difficulty level of the problem is changed or revising the question.

An example of participant answer data in SWB operation listed in Table 2. Reevaluation of the problem weight is shown in Figure 7. Weight value is used in the prediction for the next question. The performance score calculation of these participants is ∑ (2.22, 4.83, 5.1, 0, 2.08, 4.25, 4.41, 0, 2.67, 3.88) = 29.45. The process of

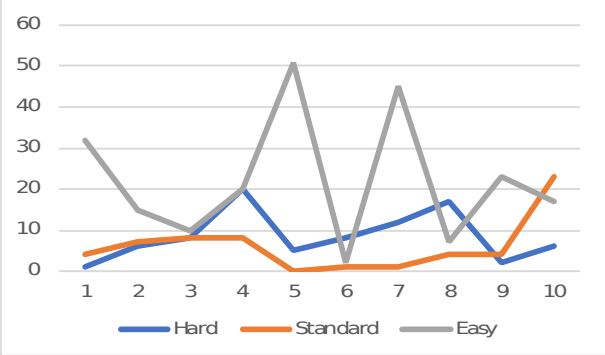

Figure 5. Correct answer level based on a difficulty level on each question

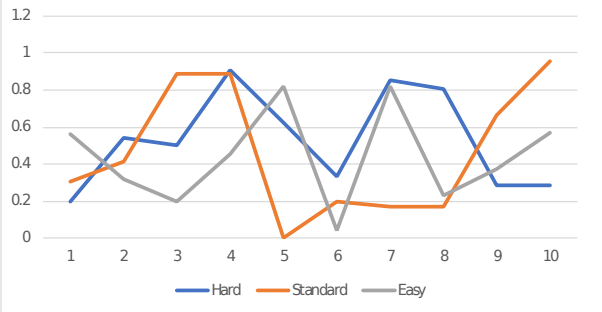

Figure 6. Probability answer correctly in each difficulty level of the question

Table 2. An example case of participant testing in SWB

#question	Mode	Answer
4	Standard	Correct
6	Hard	Correct
1	Hard	Correct
7	Hard	Wrong
10	Standard	Correct
5	Hard	Correct
2	Hard	Correct
3	Hard	Wrong
9	Standard	Correct
8	Hard	Correct

the question selection with example 5 initial questions as follows:

Question #4, standard level, correct answer. P(True) = 1, P(False) = 0

- Question #6, P(Hard) = 1.61, P(Standard) = 0.72, P(Easy) = 0.1. Selected questions are hard level, correct answer. P(True) = 1, P(False) = 0
- Question #1, P(Hard) = 1.02, P(Standard) = 1.04, P(Easy) = 0.77. Selected questions are hard level, but the standard level has the highest chance, correct answer. P(True)=1, P(False)=0
- Question #7, P(Hard) = 3.24, P(Standard) = 0.61, P(Easy) = 0.71. Selected questions are hard level, wrong answer. P(True) = 0.75, P(False) = 0.25

	1			2			3			4			5		
	W	Cor.	Wro.	W	Cor.	Wro.	W	Cor.	Wro.	W	Cor.	Wro.	W	Cor.	Wro.
Hard	5.1	0.2	0.8	4.41	0.55	0.45	4.5	0.5	0.5	3.68	0.91	0.09	4.25	0.63	0.38
Standard	3.38	0.31	0.69	3.18	0.41	0.59	2.22	0.89	0.11	2.22	0.89	0.11	4	0	1
Easy	1.38	0.56	0.44	1.86	0.32	0.68	2.1	0.2	0.8	1.59	0.45	0.55	0.85	0.82	0.18

6			7			8			9			10		
W	Cor.	Wro.	W	Cor.	Wro.	W	Cor.	Wro.	W	Cor.	Wro.	W	Cor.	Wro.
4.83	0.33	0.67	3.79	0.86	0.14	3.88	0.81	0.19	4.93	0.29	0.71	4.93	0.29	0.71
3.6	0.2	0.8	3.67	0.17	0.83	3.67	0.17	0.83	2.67	0.67	0.33	2.08	0.96	0.04
2.41	0.04	0.96	0.86	0.82	0.18	2.03	0.23	0.77	1.76	0.37	0.63	1.37	0.57	0.43

Figure 7. Weight Reevaluation with correct and wrong probability in each question with level of difficulty

- Question #10, P(Hard) = 0.18, P(Standard) = 1.48, P(Easy) = 0.43. Selected standard level questions, correct answer. P (True) = 0.8, P (False) = 0.2
- And so on until the number of correct answers is 6 out of 8 predicted questions that been executed (2 other questions do not follow the predicted questions with a balanced distribution between right and wrong answers).

The comparison with MST was conducted from a specific selected course for convenience reasons. We selected 24 students that have passed the course with A and A- grade regardless of the year and split randomly, 12 took MST and 12 took the proposed method (Selective Weight Bayesian). Both tests using the same questions material, multiple-choices and the time limit for the test is 120 minutes. We assume that the students could answer the questions because they already passed the course with grade A or A- regardless of the year. We also record the time taken by students to finish the exam. The deviation is based on the farthest score from the average value. The result is stated in Table 3. The proposed technique (SWB) has a better minimum time for the student to finish the test with an acceptable score than sequential one, so it is promising to be implemented in CAT system as [1]-[13].

The accuracy was measured based on the choice of questions at the maximum level of the examinee alleged ability by the system, compared to the correct answer from participants. It obtained an average accuracy of 75% compared to 33% sequentially. In this random model, the questions offered can be adjusted according to the ability of the participants to answer so that the system can succeed in predicting the maximum ability to answer from the examinee.

IV. Conclusion

The random model using weighted Bayesian to select a question that more likely the student can answer correctly gives a shorter time for the student to finish the exam compared to the sequential model. The system has the ability to revise the position of the problem based on the evaluation of the pattern of answers from all test participants.

Table 3. Comparison of SWB and MST

No	Description	MST	SWB
1	Average Exam Mark	7.7	7.1
2	Minimum time to finish (minutes)	52	17
3	Deviation	±0.6	±0.4

Acknowledgment

This research was supported by the LP3M unit, Universitas Maritim Raja Ali Haji with internal research funding grants, Laboratory-based research scheme.

Bibliography

[1] R. Mitkov, L. A. Ha, and N. Karamanis," A Computer-Aided Environment for Generating Multiple-Choice Test Items," *Natural Language Engineering*, vol. 12, no. 2, pp. 177-197, 2006.

[2] S.-A. A. Jin, "The Effects of Incorporating a Virtual Agent in a Computer-Aided Test Designed for Stress Management Education: The Mediating Role of Enjoyment," *Computers In Human Behavior*, vol. 26, no. 3, pp. 443-451, 2010.

[3] D. M. Williamson, R. J. Mislevy, and I. I. Bejar, *Automated Scoring of Complex Tasks in Computer-Based Testing*. New Jersey: Lawrence Erlbaum Associates, 2006.

[4] B. J. Mason, M. Patry, and D. J. Bernstein," An Examination of The Equivalence Between Non-Adaptive Computer-Based and Traditional Testing," *Journal of Educational Computing Re search*, vol. 24, no. 1, pp. 29-39, 2001.

[5] H. Jeong, "A comparative study of scores on computer-based tests and paper-based tests," *Behaviour & Information Technology*, vol. 22, no. 4, pp. 410-422, 2014.

[6] L. Wan and G. A. Henley, "Measurement Properties of Two Innovative Item Formats in a Computer-Based Test," *Applied Measurement in Education*, vol. 25, no. 1, pp. 58-78, 2012.

[7] D. Bengs, U. Brefeld, and U. Kröhne, "Adaptive Item Selection Under Matroid Constraints," *Journal of Computerized Adaptive Testing*, vol. 6, no. 2, pp. 15-36, 2018.

[8] A. Huebner, M. D. Finkelman, and A. Weissman, "Factors Affecting the Classification Accuracy and Average Length of a Variable-Length Cognitive Diagnostic Computerized Test," *Journal of Computerized Adaptive Testing*, vol. 6, no. 1, pp. 1-14, 2018.

[9] N. V. Buuren and T. H. J. M. Eggen, "Latent-Class-Based Item Selection for Computerized Adaptive Progress Test," *Journal of Computerized Adaptive Testing*, vol. 5, no. 2, pp. 22-43, 2017.

[10] B. G. Brossman and R. A. Guille, "A Comparison of Multi-Stage and Linear Test Designs for Medium-Size Licensure and Certification Examinations," *Journal of Computerized Adaptive Testing*, vol. 2, no. 2, pp. 18-36, 2014.

[11] D. I. Belov, "Detecting Item Preknowledge in Computerized Adaptive Testing Using Information Theory and Combinatorial Optimization," *Journal of Computerized Adaptive Testing*, vol. 2, no. 3, pp. 37-58, 2014.

[12] D. I. Belov, "Uniform Test Assembly: Concepts, Problems, Solvers, and Applications for Adaptive Testing," *Journal of Computerized Adaptive Testing*, vol. 5, no. 1, pp. 1-21, 2017.

[13] S. Boonsathorn, "Computer-based Test of English Competence (CB TEC) for EFL Advanced Learners: A New Format of C-Test," in *15th International Conference on Information Technology Based Higher Education and Training (ITHET)*, Istanbul, Turkey, 2016, pp. 1-5.

[14] Y. A. A. Hakami, A. R. B. C. Hussin, and H. M. Dahlan, "Technology Acceptance for CBT in Secondary Schools of Saudi Arabia," in *5th International Conference on Intelligent Systems, Modelling and Simulation*, Langkawi, Malaysia, 2014, pp. 804-807.

[15] A. Jamaluddin, D. Harjunowibowo, M. A. Rochim, F. Mahadmadi, H. B. Kakanita, and P. W. Laksono, "Implementation of RFID on Computer Based Test (RF-CBT) System," in *Proceedings of the Joint International Conference on Electric Vehicular Technology and Industrial, Mechanical, Electrical and Chemical Engineering (ICEVT & IMECE)*, Surakarta, Indonesia, 2015, pp. 153-156.

[16] M. S. B. A. Malek, M. A. B. Ahmadon, S. Yamaguchi, and B. B. Gupta, "Implementation of Parallel Model Checking for Computer-Based Test Security Design," in *7th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2016, pp. 258-263.