Journal of Informatics and Web Engineering

Vol. 4 No. 3 (October 2025) eISSN: 2821-370X

Impact of Data Quality Types on Computational Time in Data Source Selection Using Ant Colony Optimization

Nor Amalina Mohd Sabri[1*], Abd Samad Hasan Basari[2], Nurul Akmar Emran[3]

[1,2]Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Tun Hussein Onn Malaysia, Persiaran Tun Dr. Ismail, 86400 Parit Raja, Johor Darul Ta'zim, Malaysia

[3]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

**corresponding author: (noramalina@uthm.edu.my; ORCiD: 0009-0006-4461-5906)*

Abstract - Data quality varies dramatically from source to source, even within the same domain. Given these challenges, data source selection has emerged as a crucial step in information integration. It demands efficient and scalable approaches that can handle massive data volumes while ensuring the quality of results. Adapting the ACO algorithm to solve the data sources selection problems may lead to inconsistent computational time if the data sources provided are vary in quality. These challenges bring the issues of time consuming in selecting the required data sources. However, how much the computational time needed in solving the data sources selection is depending on the type of data quality. Hence, in this article, the impact of quality type of data towards computational time is examined in solving the data sources selection problems. For the methodology used, there are five steps need to be followed which are first collect data set, second import the data sources to the data sources selection model, third implement the ACO algorithm, fourth obtain the computational time and lastly compare the results. The experiment shows that low-quality data set achieve higher computational time compared to the high-quality data set which achieve the minimum computational time by 3.38 % faster. The results obtained in this experiment shown that the quality type of data has given an impact to the computational time of ACO algorithm. The results also clearly show the contribution of high-quality data set in minimizing computational time in the selection process. The validation on quality type of data with computational time is to clarify the importance of selecting a good quality data to save the computational time.

Keywords— Data Quality, Data Source Selection, Ant Colony Optimization, High-quality, Low-quality

Received: 11 May 2025; Accepted: 12 August 2025; Published: 16 October 2025

1. INTRODUCTION

The digital world has experienced an explosion in the volume and variety of data, ushering in the era of big data and open data [1]. This phenomenon is characterized by an unprecedented growth in the number of data sources,

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2025.4.3.24
© Universiti Telekom Sdn Bhd.
Published by MMU Press. URL: https://journals.mmupress.com/jiwe

potentially scaling to tens of millions, far exceeding traditional information integration scenarios [1]. Data is now considered as vital as “new oil,” underpinning data-driven decision-making in diverse fields, from public health (as seen during the COVID-19 pandemic) to business operations [2].

However, this massive influx of data comes with new features and challenges. Data quality varies dramatically from source to source, even within the same domain [1], [3]. This heterogeneity includes differing conceptual, contextual, and typographical representations [1] and even the ambiguous nature of language in fields like healthcare contributes to data quality issues [4]. There's a high probability of overlapping information among numerous data sources, partly because some sources may copy data from others [1].

Given these challenges, data source selection has emerged as a crucial step in information integration [1], [3]. It demands efficient and scalable approaches that can handle massive data volumes while ensuring the quality of results [3]. The goal of source selection is to optimize decision quality by using high-quality (HQ) data, as poor data can lead to erroneous results and financial losses [2].

The data source selection problem is generally NP-hard [3], meaning exact solutions are computationally intensive for large data sets. Therefore, research focuses on developing efficient and scalable approximate algorithms. Therefore, the Ant Colony Optimization (ACO) algorithm is utilized to solve the problems.

Adapting the ACO algorithm to solve the data sources selection problems may lead to inconsistent computational time if the data sources provided vary in quality. These challenges bring the issues of time consuming in selecting the required data sources. However, how much the computational time needed in solving the data sources selection is depends on the type of data quality. Hence, in this article, the impact of quality type of data towards computational time is examine in solving the data sources selection problems.

2. LITERATURE REVIEW

The quality of data intensely impacts data-driven decision-making across various sectors [5] and become essential part for efficient decision-making [6]. HQ data is needed in data analytic to make the data become interpretable and more trustworthy, then will lead to meaningful and accurate decisions-making [5]. Other than that, HQ data is important for modern application like artificial intelligence (AI) and for decision-making processes across various industries [7], [8]. HQ is defined as which data meets the needs and requirements setup by a business or organization and is accepted for usage in required that business processes [9], [6]. There are key dimensions of HQ data include accuracy, completeness, and consistency which often highlighted to define the degree of data quality [7]. However, the demand for data in various applications are increasing rapidly where brings significant challenges towards its quality [7]. Meanwhile, poor data quality, frequently referred to as “dirty data,” can affect the effectiveness of optimization models and lead to negative results or outcomes [9], [10], [11].

Furthermore, the impact of poor data quality in decision-making process may lead to erroneous results. Poor data quality also can lead to incorrect results and outcomes, especially in optimization models that involve on large data sets which used for training and testing [2], [4]. However, if the data used to train optimization models, are collected by incapable algorithms, or contain errors, the prediction results can be masked, therefore making it difficult to differentiate between correct and incorrect outcomes [4]. Other than that, the critical impact of data quality problems creates to an increasing amount of unusable data [12]. Furthermore, poor data quality can make running time increased for algorithms because of the uncertainty, irrelevant and redundant features also complexity in that data [11], [13].

To prevent these impacts, data source selection is a crucial step in demanding efficient and scalable algorithms [1]. Optimization is a pervasive technique that has been applied across numerous fields including data source selections. Which is fundamentally a process or technique of improving or enhancing a model, system, or method to complete a more optimal outcome [14]. This often involves a thoughtful effort to increase desirable characteristics and or minimize undesirable ones [15]

In this big data era, to make a subset selection of data sources from massive and various selections is a key to optimization problem. This involves balancing the efficiency by scaling to large source amounts and effectiveness in terms of data quality and source overlapping [3]. However, to access all the data sources is costly and quite impossible, Therefore, by selecting a small amount of HQ data sources and relevant data sources is important to achieve better efficiency and effectiveness [1].

The optimization technique is a viable method in solving selection problem since it seeks to maximize output and minimize computational time, and at the same time able to maintain the proper leverage between quality and performance of the results achieved. Hence, the optimization algorithm is utilized to enhance the selection process efficiently. However, the selection process turns into a challenging problem because of the large volume of data. But the selection process can become stable when the performance is improves and computing cost decreases [16].

Optimization refers to the process of finding the optimal variable for a specific problem in by minimizing or maximizing a well-designed objective function [17]. In the big data era, optimization is critically needed where effective decision-making hinges on the quality of data and the efficiency of processing large data sets [2]. Moreover, many complex optimization problems, such as data source selection and cloud service composition, are considered NP-hard [18]. This classification shows that finding optimal solutions for highlighted problems is computationally very time-consuming, especially when the problem scale increases [19].

ACO is a metaheuristic algorithm for optimization that inspired by the social behaviour of ants while searching for the shortest paths to find food sources [20]. ACO algorithms are effective for solving NP-hard combinatorial optimization problems, including those in cloud computing area to solve task scheduling, resource allocation, and service composition problems [19]. ACO is a swarm intelligence algorithm adopted by the foraging behaviour of real ants [21]. It is highlighted as powerful computational tool were designed to solve combinatorial optimization problems, where the objective is to find optimal solutions within a graph or network [17]. ACO is also distributed cluster optimization algorithm known for its speed in finding global optimal solutions [21]. Moreover, ACO is claimed as robust, easily adaptable, uses a probabilistic decision-making process for efficient search space exploration, and has positive and negative feedback mechanisms for pheromone reinforcement and evaporation [17]. Its cooperative behaviour allows ants to indirectly communicate and able to find optimal solutions [17]. Initially, ACO was used to solve problems like the Traveling Salesman Problem (TSP) and has been effectively applied to various difficult combinatorial optimization problems such as feature selection problem [20].

Additionally, ACO can search locally because of the ACO's stochastic component that can effectively searches the space where the problem of getting stuck in a local minimum can be avoided. Furthermore, ACO is thought to be an intelligent agent because of its characteristics to have high degree of self-organization and its ability for complex problem performance. It also motivates a lot of researchers to create a new finding for computer science problem optimization [22].

The use of ACO in optimization has been the subject of various research, but the impact of different data quality types on computation performance has received less attention. Furthermore, previous research assesses algorithm efficiency in terms of accuracy or solution quality rather than methodically examining the ways in which various aspects of data quality affect computation time. The current study, which attempts to investigate the connection between HQ data types and ACO's computational efficiency in the context of data source selection, is motivated by this gap.

3. RESEARCH METHODOLOGY

The methodology adopted to this work are shown in Figure 1.

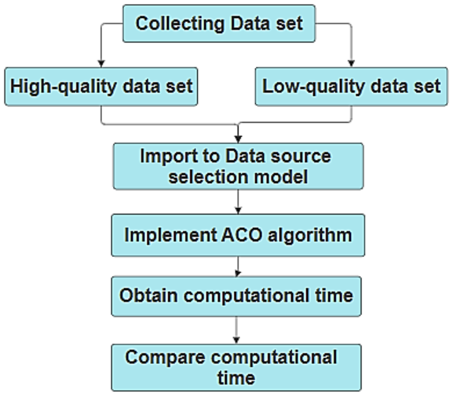
Collecting Data set
High-quality data set
Low-quality data set
Import to Data source selection model
Implement ACO algorithm
Obtain computational time
Compare computational time

Figure 1. Research Methodology

To validate the impact of quality type of data and computational time in solving data sources selection problem using ACO algorithm, there are five steps need to be followed as shown in Figure 1. First step is collecting data set; the data set is collected originally from index.okfn.org open government data portal ranking list. This portal has made a ranking of open government portal all around the world by their own required criteria. The criteria are divided based on the elements/categories that should be fulfilled by the government in sharing their data to citizen such as data related to Government budget, National Laws and Government spending. Those open government data did not fulfil the criteria in terms of completeness of data shared will have less score and will be ranking according to the score value. This data set is used to create two types of data sets which are HQ data set and low-quality (LQ) data set. The data sets are sorted by the top 30 and bottom 30 ranking lists, which establishes the degree of quality score data for each set. While the lowest 30 ranking list is referred to as a LQ data set, the top 30 ranking list is classified as a HQ data set. According to the open government data portal ranking list the HQ data set is defined as a completed data sets that fulfilled all the quality criteria required. Meanwhile, the LQ data set is not fulfilling the required quality criteria which make the data set has the missing values.

In second step, both data sets are imported to the data source selection model separately. In the data set itself has the quality score which will be selected by the ACO algorithm. Higher quality score represents HQ data sources in the HQ data set, while the LQ data set has inconsistent quality score data sources. Inconsistent quality scores data sources exist because of the incompleteness of every data source in fulfilling the quality criteria required by the open data government portal.

Third step in this research work is to implement the ACO algorithm into the model. The implementation of ACO algorithm is to find the most HQ data sources in the data set. With the performance of ACO in finding the optimal solution in solving selection problem, this algorithm able to sort the data sources from the most HQ to lesser quality of data sources. The experiment is done separately for HQ data set and LQ data set; the experiment is run for 20 times with different number of data sources start from 7 data sources until 452 numbers of data sources. This repetition is made to have precise and consistent results. From the experiment, the computational time of ACO algorithm to complete the searching of optimal solution are recorded. There are two sets of results which are for HQ data set result and LQ data set result.

Finally, the result from both data sets which are HQ data set and LQ data set are compared. The comparison of the results aims to validate the performance of ACO algorithm in selecting data sources from high quality data set and LQ data set. This final step help to conclude the objective of this research towards the impact of quality type of data on computational time generated by ACO algorithm in solving the data sources selection problem.

4. RESULTS AND DISCUSSIONS

In this article, the impact of different quality types of data on computational time during data source selection using ACO is validated by comparing the results of computational time for every type of quality data which are HQ data set and LQ data set that have been generated by ACO in selecting the data sources.

Two variables which are the number of data sources and computational time are utilized to validate these two data sets. Up until 452 data sources, the number of data sources is arranged in ascending order. From the start of the algorithm's execution to its completion, the computing time is tracked. The computational time and number of data sources for HQ data set are recorded in Table 1.

The computational time is recorded from 0.4227 seconds until 327.1541 seconds for HQ data set. The computational time is increasing as the number of data sources increase. The average computational time for HQ data set is 52.4847. as depicted in Table 1.

The computational time obtained for LQ data set is between 0.4121 seconds and 397.4326 seconds, increasingly obtained when the number of data sources increased. The average computational time for LQ data set is 60.2018 seconds. The graph of comparison results for HQ and LQ computational time is shown in Figure 2.

For the total of 452 data sources, HQ data set compared to LQ data set is 3.38 % faster. Figure 2 shows that LQ data set achieve higher computational time for 452 data sources compared to the HQ data set which achieve the minimum computational time. The graph in Figure 2 shows that the computational time of LQ data set begin to have bigger difference from HQ data set on experiment number 18 with 377 number of data sources for over than 15 seconds. Meanwhile, the computational time graph for HQ data set is increase slightly consistent.

Table 1. HQ and LQ Computational Time

No of experiment	Number of data sources	HQ Computational time (s)	LQ Computational time (s)
1	7	0.4227	0.4121
2	17	0.4618	0.4612
3	24	0.5832	0.5925
4	32	0.6723	0.6772
5	47	2.1396	2.1086
6	69	2.9560	2.9266
7	84	3.6658	3.7294
8	107	5.2545	5.2635
9	122	6.6796	6.7032
10	137	8.6055	8.5520
11	159	12.2535	12.3337
12	174	15.1327	15.2062
13	197	20.7681	20.1200
14	227	29.3828	29.3975
15	272	51.8735	51.2694
16	302	69.5317	76.6226
17	332	95.8923	101.7464
18	377	163.8332	178.8498
19	407	196.0389	264.6203
20	452	327.1541	397.4326
Average		52.4847	60.2018

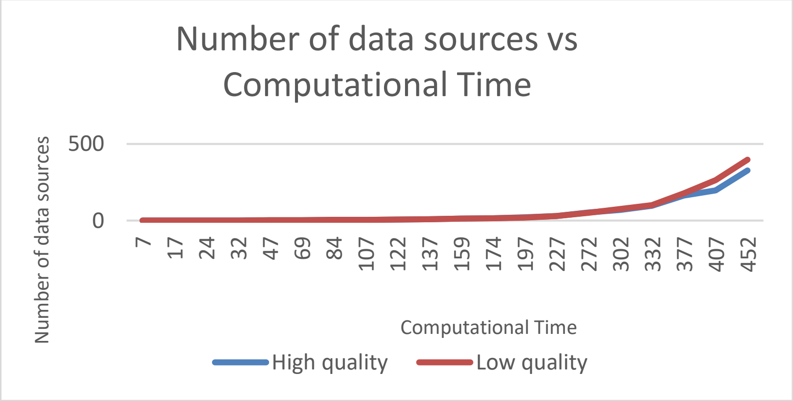

Figure 2. Comparison Results

From these results, we can validate that the quality scores or quality type of data affect the performance of ACO algorithm in terms of computational time in solving the data sources selection problems. This is because ant need shorter time to traverse the graph around HQ data set due to small differences of quality score for every data source. Furthermore, the HQ data set has less missing quality scores as compared to low quality data set which contributes to consistent quality scores. Meanwhile the quality score of LQ data set is inconsistent value, that force the ACO algorithm to use longer time to traverse the graph or path to select the optimal data sources.

5. CONCLUSION

In this article, the impact of quality type of data and the computational time of ACO algorithm in solving the data sources selection problem is validated. The relationship between number of data sources and computational time taken by the ACO algorithm are obtained from the experiment in the data sources selection model. The experiment is run repeatedly by increasing the number of data sources up to 452 number of data sources and the computational time is recorded. The experiment is done separately for HQ and LQ data set. The results obtained in this experiment shown that the quality type of data has given an impact to the computational time of ACO algorithm. The results also clearly show the contribution of HQ data set in minimizing computational time taken by the ACO algorithm in the selection process. The validation on quality type of data with computational time is to clarify the importance of selecting a good quality data to save the computational time and the obtained result shows the HQ data set is 3.38 % faster than LQ data set.

However, a statistical analysis was conducted to examine whether data quality (HQ vs. low) has a significant effect on computational time in solving the data source selection problem. The Shapiro-Wilk normality test revealed that the computational time data for both HQ and LQ data sets were not normally distributed ($p < 0.001$). Despite this, Levene’s test indicated that the variances between the two groups were equal ($p = 0.778$). A Welch’s t-test showed no statistically significant difference in the mean computational time between HQ and LQ data ($t = -0.273$, $p = 0.787$). This result was further supported by the Mann–Whitney U test, a non-parametric alternative, which also showed no significant difference in the distribution of computational times ($U = 197.0$, $p = 0.946$). Therefore, it can be concluded that the type of data quality does not have a significant impact on the computational time required for the data source selection process in this experiment.

ACKNOWLEDGEMENT

The authors express their gratitude to the Faculty of Computer Science and Information Technology (FSKTM) at Universiti Tun Hussein Onn Malaysia (UTHM).

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Nor Amalina Mohd Sabri: Conceptualization, Formal Analysis, Experiment, Methodology, Validation, Visualization, Writing – Original Draft Preparation;
Abd Samad Hasan Basari: Supervision, Writing – Review & Editing;
Nurul Akmar Emran: Supervision, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

REFERENCES

[1] Y. Lin, H. Wang, J. Li, and H. Gao, "Data source selection for information integration in big data era," *Information Sciences*, vol. 479, pp. 197–213, 2019.

[2] J. Wang *et al.*, "Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality," *Journal of the Knowledge Economy*, vol. 15, pp. 1159–1178, 2024, doi: 10.1007/s13132-022-01096-6.

[3] Y. Lin, H. Wang, S. Zhang, J. Li, and H. Gao, "Efficient quality-driven source selection from massive data sources," *The Journal of Systems and Software*, vol. 118, pp. 221–233, 2016, doi: 10.1016/j.jss.2016.05.026.

[4] F.A. Bernardi, D. Alves, N. Crepaldi, D.B. Yamada, V.C. Lima, and R. Rijo, "Data Quality in Health Research: Integrative Literature Review," *J Med Internet Res*, vol. 25, pp. e41446, 2023, doi: 10.2196/41446.

[5] L. Ehrlinger and W. Wöß, "A Systematic Review of Data Quality Measurement and Monitoring Tools," *Frontiers in Big Data*, vol. 5, Art. no. 850611, 2022, doi: 10.3389/fdata.2022.850611.

[6] T. Peixoto *et al.*, "Data Quality Assessment in Smart Manufacturing: A Review," *Systems*, vol. 13, no. 243, pp. 1–28, Mar. 2025, doi: 10.3390/systems13040243.

[7] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, "The effects of data quality on machine learning performance on tabular data," *Information Systems*, vol. 132, p. 102549, Mar. 2025, doi: 10.1016/j.is.2025.102549.

[8] Matoni, A. Kesper, and G. Taentzer, "How to Define the Quality of Data? A Feature-Based Literature Survey," *arXiv preprint arXiv:2504.01491* 2025.

[9] F. Ridzuan, and W. M. N. W. Zainon, "A Review on Data Quality Dimensions for Big Data," *Procedia Computer Science*, vol. 234, pp. 341-348, 2024, doi: 10.1016/j.procs.2024.03.008.

[10] H. Cho, and S. Lee, "Data Quality Measures and Efficient Evaluation Algorithms for Large-Scale High-Dimensional Data," *Appl. Sci.*, vol. 11, no. 2, pp. 472, Jan. 2021, doi: 10.3390/app11020472.

[11] Z. Qi, H. Wang, J. Li, and H. Gao, "Impacts of Dirty Data: an Experimental Evaluation," arXiv:1803.06071v2, Mar. 2018.

[12] O. Ozonze, P.J. Scott, and A.A. Hopgood, "Automating Electronic Health Record Data Quality Assessment," *Journal of Medical Systems*, vol. 47, Art. no. 23, 2023, doi: 10.1007/s10916-022-01892-2.

[13] L.G. Fahad et al., "Ant Colony Optimization-Based Streaming Feature Selection: An Application to the Medical Image Diagnosis," Sci. Program., vol. 2020, Article ID 1064934, pp. 1–10, Oct. 2020, doi: 10.1155/2020/1064934.

[14] M. R. Abdurrahman, H. Al-Aziz, F.A. Zayn, M.A. Purnomo, and H.A. Santoso, "Development of Robot Feature for Stunting Analysis Using Long-Short Term Memory (LSTM) Algorithm," *J. Informatics Web Eng.*, vol. 3, no. 3, pp. 165–175, Oct. 2024, doi: 10.33093/jiwe.2024.3.3.10.

[15] P.-W. Chin, K.-W. Ng, and N. Palanichamy, "Plant Disease Detection and Classification Using Deep Learning Methods: A Comparison Study," *J. Informatics Web Eng.*, vol. 3, no. 1, pp. 156–168, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.10.

[16] N.A.M. Sabri, N.A. Emran, N. Harum, "Open Government Data (OGD) portals selection using Ant Colony Optimization (ACO) Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, vol 9, pp. 6555-6562, 2020.

[17] M.A. Awadallah *et al.*, "Multi-objective Ant Colony Optimization: Review," *Archives of Computational Methods in Engineering*, vol. 32, pp. 995–1037, 2025, doi: 10.1007/s11831-024-10178-4.

[18] F. Dahan, "An Effective Multi-Agent Ant Colony Optimization Algorithm for QoS-Aware Cloud Service Composition," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3052907.

[19] N. Sharma, Sonal, and P. Garg, "Ant colony based optimization model for QoS-based task scheduling in cloud computing environment," *Measurement: Sensors*, vol. 24, p. 100531, 2022, doi: 10.1016/j.measen.2022.100531.

[20] H.R. Kanan and K. Faez, "An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system," *Applied Mathematics and Computation*, vol. 205, pp. 716-725, 2008, doi: 10.1016/j.amc.2008.05.115.

[21] Z. Zhang, J. Li, and N. Xu, "Robust optimization based on ant colony optimization in the data transmission path selection of WSNs," *Neural Computing and Applications*, vol. 33, pp. 17119–17130, 2021, doi: 10.1007/s00521-021-06303-0.

[22] N.A.M. Sabri, N. A. Emran, N. Abdullah, "Quality-Based Open Data Source Selection Using Ant Colony Optimization (ACO) Algorithm," *International Journal on Emerging Technologies*, vol 11, pp. 1164-1168, 2020.

BIOGRAPHIES OF AUTHORS

	Nor Amalina Mohd Sabri is an academician at Universiti Tun Hussein Onn Malaysia. She obtained her bachelor's degree in information technology from Universiti Tun Hussein Onn Malaysia in 2012 and then Master of Science in Information Technology from Universiti Teknikal Malaysia Melaka in 2016. In 2022, she was awarded a Doctor of Philosophy (PhD) by Universiti Teknikal Malaysia Melaka, Malaysia with her doctoral research concentrating on optimization using evolutionary algorithm. She can be contacted at email: noramalina@uthm.edu.my.
	Abd Samad Hasan Basari is an academician at Universiti Tun Hussein Onn Malaysia. He obtained his Bachelor of Science (Hons.) degree in Mathematics from Universiti Kebangsaan Malaysia in 1998 and then Master of Information Technology in Education (Computational Modelling) from Universiti Teknologi Malaysia in 2002. In 2009, he was awarded a Doctor of Philosophy (PhD) by Universiti Teknikal Malaysia Melaka, Malaysia with his doctoral research concentrating on Maintenance Modelling with a specific focus on incomplete data. He can be contacted at email: abdsamad@uthm.edu.my.
	Nurul Akmar Emran is an Associate Professor in Software Engineering Department. She begins her career in academic in 2002, where she joined Kolej Universiti Teknikal Malaysia Melaka in 2003 as a tutor at the Faculty of Information and Communication Technology. In 2004, she was appointed as a lecturer; in 2011 she becomes a senior lecturer after successfully completing her PhD at the University of Manchester. In 2017, she was promoted as Associate Professor. She can be contacted at email: nurulakmar@utem.edu.my.