

Taguchi's T-method with nearest integer-based binary bat algorithm for prediction

Zulkifli Marlah Marlan¹, Khairur Rijal Jamaludin¹, Faizir Ramlie¹, Nolia Harudin²

¹Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

²Department of Mechanical, College of Engineering, Universiti Tenaga Nasional, Kajang, Malaysia

Article Info

Article history:

Received Mar 28, 2022

Revised May 28, 2022

Accepted Jun 10, 2022

Keywords:

Binary bat algorithm

Feature selection

Mahalanobis-Taguchi system

Nearest integer discretization

Taguchi's T-method

ABSTRACT

Taguchi's T-method is a new prediction technique under the Mahalanobis-Taguchi system to predict unknown output or future states based on available historical information. Conventionally, in optimizing the T-method prediction accuracy, Taguchi's orthogonal array is utilized to determine a subset of significant features to be used in formulating the optimal prediction model. This, however, resulted in a sub-optimal prediction accuracy due to its fixed and limited feature combination offered for evaluation and lack of higher-order feature interaction. In this paper, a swarm-based binary bat optimization algorithm with a nearest integer discretization approach is integrated with the Taguchi's T-method. A comparative study is conducted by comparing the performance of the proposed method against the conventional approach using mean absolute error as the performance measure on four benchmark case studies. The results from experimental studies show a significant improvement in the T-method prediction accuracy. A reduction in the total number of features results in a less complex model. Based on the general observation, the nearest integer-based binary bat algorithm successfully optimized the selection of significant features due to recursive and repetitive searchability, in addition to its adaptive element in response to the current best solution in guiding the search process towards optimality.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Zulkifli Marlah Marlan

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia

Jalan Sultan Yahya Petra (Jalan Semarak), 54100 Kuala Lumpur, Malaysia

Email: zulkifli49@graduate.utm.my

1. INTRODUCTION

A predictive analytic technique is an essential tool in obtaining information about the future state or unknown outcome by analyzing current or historical data. The growth of big data, the availability of efficient software and hardware for data processing, and a competitive and agile business environment have motivated the development of various prediction approaches [1]. Some examples include Bayesian network, artificial neural network, multiple linear regression, logistic regression, and machine learning algorithms such as random forest [2], [3]. One of the relatively new predictive modeling technique that is capable of producing an accurate outcome is the Taguchi's T-method (T-method), which was introduced by Dr. Genichi Taguchi under the Mahalanobis-Taguchi system (MTS) [4]. Out of many methods under MTS architecture, the T-method is designed explicitly for solving prediction-based problems involving multivariate information. Nevertheless, the T-method is also capable of performing classification tasks with the aid of interval values computed according to classified groups or items. Fundamentally, the T-method is developed based on the combination of the regression principle and Taguchi's robust quality engineering elements in formulating a

predictive model. The blend of mathematical-statistical theory such as linear regression and weighted average model with unit space element and orthogonal array (OA) experimental design has become the significant differentiation factor that distinguished the T-method from other available predictive techniques.

As a multivariate prediction technique, the T-method predictive model is formulated with multiple input features to predict the output. As a consequence, the prediction model becomes complex as the dimensionality of data gets higher, and in many situations, not all input features are significant and relevant towards the prediction outcome [5]. Conventionally, in optimizing the T-method predictive model, an OA design was employed for determining a subset of significant input features based on the maximum value of the integrated estimate signal-to-noise ratio as the performance measure. Researchers [6] and [7] highlighted that the application of OA in performing feature selection is insufficient and yields a sub-optimal solution. It is mainly due to the OA's fixed and limited variable combination, limiting the possibility of finding the optimal solution [8]. In addition, Kim *et al.* [9] stressed that the OA lacks exploiting higher-order interactions (mixture) between variables, leading to sub-optimality of prediction accuracy. In response to these concerns, this paper proposed the utilization of a swarm-based binary bat optimization algorithm (BBA) with the nearest integer discretization approach as an alternative to the OA.

The utilization of metaheuristic algorithms such as the BBA for feature selection optimization problems is not new. Many studies reported the success of metaheuristic algorithms in solving combinatorial problems, as demonstrated in [10], [11]. There are many types of metaheuristic algorithms available for consideration that are categorized into four according to their search behaviors, which are evolution-based, swarm intelligence-based, physics-based, and human-related algorithms [12]. The BBA belongs to the swarm intelligence-based along with particle swarm optimization (PSO), ant colony optimization, honey bee swarm optimization algorithm, cuckoo search optimization, and many others. Specifically, in optimizing the T-method prediction accuracy, Harudin *et al.* [13] successfully employed the artificial bee colony algorithm for feature selection optimization. The result shows an improvement in prediction accuracy as compared to the conventional OA approach. In a different study, Harudin *et al.* [14] utilized a modified artificial bee colony algorithm with a binary bitwise operator as the feature selection approach, and the outcome recorded an enhancement in prediction accuracy. These studies have shown practicality in employing metaheuristic algorithms as the T-method feature selection optimization.

2. METHOD

The development of an optimal T-method prediction model consists of two main phases. The first phase focuses on the development of the basic prediction model, while the second phase concentrates on the optimization of the formulated model through a feature selection process. This paper focused on the second phase to replace the conventional approach using Taguchi's OA with Nearest Integer-based Binary Bat algorithm.

2.1. Development of the basic T-method prediction model

Prior to optimizing the T-method prediction model using the proposed approach, the basic prediction model must be established. The development of the basic T-method predictive model involved the determination of two important model parameters known as a proportional coefficient, β and signal-to-noise ratio (SNR), η . In estimating the model's parameters, the raw data is first transformed into signal data through a normalization process using the average of unit space data. Unit space is a concept emphasized in the Mahalanobis-Taguchi system that represents a homogeneous population against the target group [15]. As such, a subset of homogeneous output data located in a densely populated region is selected, and the average value is computed for the output and respective input features. Normalization of signal data is performed by subtracting the average value from raw data, and the unit space data is discarded from signal data for further computation activity. As a result of normalization, the prediction model in the form of a linear regression line has zero intercept value (through the origin). Theoretically, the T-method predictive model is represented as an integrated estimate output value that measures the predicted outcome, as shown in (1). It was formed by combining all input features by considering the effect of the respective model parameters using a weighted average approach.

$$\text{Integrated Estimate Output Value, } \hat{M}_i = \frac{\eta_1 \times \frac{X_{i1}}{\beta_1} + \eta_2 \times \frac{X_{i2}}{\beta_2} + \dots + \eta_k \times \frac{X_{ik}}{\beta_k}}{\eta_1 + \eta_2 + \dots + \eta_k} \quad (1)$$

where X_i is the normalized signal data for respective i^{th} signal data and k is the number of features. Estimation of model's parameters performed using (2) and (4):

$$\text{Proportional coefficient, } \beta_j = \frac{M_1 X_{1j} + M_2 X_{2j} + \dots + M_l X_{lj}}{r} \quad (2)$$

where j is the input feature, M is the normalized output of signal data, l is the number of signal data, and r is the effective divider computed using (3).

$$\text{Effective divider, } r = M_1^2 + M_2^2 + \dots + M_l^2 \quad (3)$$

$$\text{SNR, } \eta_j \begin{cases} = \frac{\frac{1}{2}(S_{\beta j} - V_{ej})}{V_{ej}}; & (\text{when } S_{\beta j} > V_{ej}) \\ = 0; & (\text{when } S_{\beta j} \leq V_{ej}) \end{cases} \quad (4)$$

where V_e is the error variance obtained by (5) until (8).

$$\text{Error variance, } V_{ej} = \frac{S_{ej}}{l-1} \quad (5)$$

$$\text{Error variation, } S_{ej} = S_{Tj} - S_{\beta j} \quad (6)$$

$$\text{Total variation, } S_{Tj} = X_{11}^2 + X_{21}^2 + \dots + X_{l1}^2 \quad (7)$$

$$\text{Variation of proportional term, } S_{\beta j} = \frac{(M_1 X_{11} + M_2 X_{21} + \dots + M_l X_{l1})^2}{r} \quad (8)$$

An integrated estimate SNR (db) is computed using (9) until (15) to quantitatively represent the model's quality upon establishing the T-method prediction model and obtained the predicted value, \widehat{M} .

$$\text{Integrated estimate SNR (db), } \eta_{est} = 10 \log \frac{\frac{1}{2}(S_{\beta j} - V_{ej})}{V_{ej}} \quad (9)$$

$$\text{Linear equation, } L = M_1 \widehat{M}_1 + M_2 \widehat{M}_2 + \dots + M_l \widehat{M}_l \quad (10)$$

$$\text{Effective divider, } r = M_1^2 + M_2^2 + \dots + M_l^2 \quad (11)$$

$$\text{Total variation, } S_T = \widehat{M}_1^2 + \widehat{M}_2^2 + \dots + \widehat{M}_l^2 \quad (12)$$

$$\text{Variation of proportional term, } S_\beta = \frac{L^2}{r} \quad (13)$$

$$\text{Error variation, } S_e = S_T - S_\beta \quad (14)$$

$$\text{Error variation, } S_e = S_T - S_\beta \quad (15)$$

2.2. Taguchi's orthogonal array for feature selection

In brief, Taguchi's orthogonal array is a fractional orthogonal design with a predetermined design array proposed by Dr. Genichi Taguchi, which reduced the number of experiments run significantly, allowing for efficient sampling of multidimensional design space [16]. From the perspective of feature selection process, the utilization of a fractional orthogonal design with a predetermined design array suggested that optimality of the T-method prediction model could not be achieved using Taguchi's orthogonal array as the process of feature selection is not dynamic in evaluating numerous potential feature combinations to return the optimal (near-optimal) solution. Nevertheless, Taguchi's orthogonal array offers speed in computation and less experimental cost by utilizing a balanced predetermined design array. Specifically for feature selection optimization, a 2-level orthogonal array is used to represent the 'used' and 'not used' state of the feature in the combination. As for the T-method, the selection procedure of significant input features comprises four steps. The first step involved a suitable selection of OA design based on the number of input features. The second step involved the estimation of integrated estimate SNR (db) for each combination in the selected OA using (1) until (15). Next, the third step involved estimation of the average value of integrated estimate SNR (db) for each level or state for each feature. The last step is evaluating the average differences between levels, which can be done by developing a factorial effect chart to visualize the differences. A reduction in average integrated estimate SNR (db) from 'used' state to 'not state' of feature signify that the respective feature plays an important role in contributing to the prediction outcome.

2.3. Binary bat algorithm

The consideration of using a swarm-based metaheuristic algorithm to optimize the T-method's feature selection process is highly due to its ability to offer an optimal (near-optimal) solution. Some of its prominent traits are the ability to avoid premature solutions through exploration and exploitation of solution space strategy, stochastic and adaptive behavior in dealing with a complex problem, simplicity, ease of implementation, and operates at a reasonable computation cost [12]. Specifically, a bat algorithm is selected, which was developed based on the natural echolocation ability of microbats to identify prey, discriminate between different types of targets, and avoid obstacles [17]. Conceptually, microbats generate a very loud sound pulse with varying qualities. The echo that bounces back from the surrounding item is listened to and processed for various reasons, including hunting and navigation. Accordingly, three approximation rules are suggested in the development of the bat algorithm [18]:

- a. Bats utilize echolocation to sense distance and can distinguish between food, prey, and background barriers in a miraculous manner;
- b. Bats search for prey by flying at a random velocity, v_i at position, x_i with a fixed frequency, f_{min} , varying wavelength, and loudness, A_0 . Depending on the closeness of their target, bats may automatically modify the wavelength of their emitted pulses as well as the rate of pulse emission, $r \in [0, 1]$;
- c. Although loudness can vary in a variety of ways, the loudness assumed ranges from a maximum, A_0 to a minimum, A_{min} .

In the bat algorithm, artificial bat, i moved by updating frequency, f , velocity, v_i and position, x_i using (16), (17), and (18).

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (16)$$

$$v_i^t = v_i^{t-1} + (x_i^t - x_*)f_i \quad (17)$$

$$x_i^t = x_i^{t-1} + v_i^t \quad (18)$$

where $\beta \in [0,1]$ is a random number generated from a uniform distribution, t is the time step or iteration and x_* is the current global best solution. To further exploit the solution for optimality, a random walk is performed using (19) on the selected best solution.

$$x_{new} = x_{old} + \varepsilon \bar{A}^t \quad (19)$$

where $\varepsilon \in [0,1]$ is a random number from uniform distribution and \bar{A}^t is the average loudness, A of all bats at iteration, t . The balance between exploration on the global scale and exploitation of solutions in the targeted region of the bat algorithm is controlled by the loudness, A , and pulse emission rate, r . Both parameters are updated accordingly as iteration grows using (20) and (21):

$$A_i^{t+1} = \alpha A_i^t \quad (20)$$

$$r_i^{t+1} = r_i^0 [1 - \exp(-\gamma t)] \quad (21)$$

where α and γ are constant. The pseudocode of the bat algorithm is available in [17].

The binary version of the bat algorithm was introduced by [19] and [20] for discrete and feature selection optimization. Instead of moving in a continuous solution space, bats in the binary bat algorithm move across the corner of a hypercube space modeled as an n -dimensional Boolean lattice. In the feature selection optimization problems, a binary vector in the form of '1' and '0' is used to represent the used and not used of feature in binary-string combination. As such, the position is restricted to binary-valued, and updating the bat's position means switching between '1' and '0' values. [19] and [20] suggested that switching should be done based on the velocity of bats and proposed sigmoid-based and v-shaped transfer functions, respectively, to map the continuous velocity values into a probability value before employing a binary operator in determining the binary position of bats. For the sigmoid transfer function, continuous-valued of bat velocity v_i is transformed into probability-valued between 0 and 1 using (22). The position of the bat is then determined using a binary operator, as shown in (23). Meanwhile, the V-shaped transfer function uses (24) to transform continuous into probability value and determine the position of the bat using (25).

$$\text{Sigmoid function, } S(v_i^k(t)) = \frac{1}{1 + e^{-v_i^k(t)}} \quad (22)$$

$$x_i^k(t+1) = \begin{cases} 0; & \text{if } rand < S(v_i^k(t)) \\ 1; & \text{if } rand \geq S(v_i^k(t)) \end{cases} \quad (23)$$

$$V\text{-shaped function, } V(v_i^k(t)) = \left| \frac{2}{\pi} \arctan\left(\frac{\pi}{2} v_i^k(t)\right) \right| \quad (24)$$

$$x_i^k(t+1) = \begin{cases} (x_i^k(t))^{-1}; & \text{if } rand < V(v_i^k(t)) \\ x_i^k(t); & \text{if } rand \geq V(v_i^k(t)) \end{cases} \quad (25)$$

where $v_i^k(t)$ is the velocity of i^{th} bat for k^{th} feature at iteration t , and the $x_i^k(t+1)$ is the position i^{th} bat for k^{th} feature at iteration of $t+1$.

2.4. Proposed nearest integer-based binary bat algorithm (NIBBA)

In this paper, the nearest integer discretization approach is used in the binary bat algorithm for transforming the continuous bat's position into binary values. It was reported in [21] that this approach was first introduced in [22] for solving reactive power and voltage control optimization problems. Later [23] employed the nearest integer discretization approach for the binary bat algorithm. Theoretically, this approach involved rounding the continuous-valued bat's position to the nearest integer for each feature or dimension using (26):

$$X_i^k = \lfloor \lfloor x_i^k \bmod 2 \rfloor \rfloor \quad (26)$$

where x_i^k is the position of i^{th} bat for k^{th} feature [24]. Firstly, the continuous-valued position of the bat is divided by two. Then, the absolute value of the remainder is floored in obtaining a binary value of either 0 or 1. For instance, given the continuous-valued position of a bat is 3.45, the computational procedure begins by dividing 3.45 by 2, resulting in the remaining 1.45. The absolute value of $|1.45|$ is then floored $\lfloor 1.45 \rfloor$ resulting in the value of 1. The process is repeated for every feature or dimension until a binary string of combinations is obtained. Contrary to the sigmoid and v-shaped transfer functions approaches that depend on the bat's velocity in determining the binary position, the nearest integer method switches the position itself. This, however, is criticized as there is a possibility that the solution obtained is not in the optimal region, and the quality of the solution in the rounded point does not represent the original continuous position [21]. The pseudocode of the proposed NIBBA is as shown in Figure 1. The discretization of the continuous position of the bat using the nearest integer scheme is performed at line 4 for the initial population's position and line 13 for the new bat's position.

NIBBA Algorithm

- 1: Initialize random continuous bat population x_i ($i = 1, 2, \dots$, population size)
- 2: Initialize bat velocity, $v_i = 0$
- 3: Initialize bat frequency, f_i at x_i ,
- 4: Initialize bat pulse rate, r_i , and loudness, A_i
- 5: Transform continuous-valued position, x_i into a binary value using (26)
- 6: Evaluate fitness
- 7: Find the best: fitness, binary position, x^* , continuous position, x^{**}
- 8: **While** (iteration < Maximum iterations)
- 9: Generate new solutions:
- 10: Adjust frequency: $f_i = f_{\min} + (f_{\max} - f_{\min})\beta$
- 11: Updating velocity: $v_i^t = v_i^{t-1} + (x_i^t - x^{**})f_i$
- 12: Updating position new: $x_i^t = x_i^{t-1} + v_i^t$
- 13: **if** (rand > r_i)
- 14: Select a solution among the best solutions.
- 15: Generate a local solution around the selected best solution
- 16: Transform continuous-valued new position, x_i into a binary value using (26)
- 17: **end**
- 18: Evaluate new fitness
- 19: **if** (rand < A_i & $f(x_i) < f(x^*)$)
- 20: Accept the new solutions
- 21: Increase pulse rate, r_i and reduce loudness, A_i
- 22: **end**
- 23: Rank the bats and find the current best x^*
- 24: **end while**

Figure 1. Pseudocode of the NIBBA algorithm

3. RESULTS AND DISCUSSION

3.1. Experimental design

In determining the performance of the proposed T-method with the NIBBA algorithm, an experimental study was conducted using four benchmark datasets, as shown in Table 1. The first three datasets were obtained from the University of California at Irvine (UCI) machine learning repository [25], while the body fat dataset was obtained from the Carnegie Mellon University StatLib repository [26]. All datasets that come from different case study domains consist of multiple input features and a single output with a different number of samples. Such variability is purposely designed to allow for better performance verification of the proposed method in dealing with varying problem settings and conditions. A hold-out cross-validation is employed to the datasets, where 70% of the data will be used as train data set for the learning of the prediction model, while the remaining 30% is allocated as validation data set to verify the effectiveness of the model when tested on a newly seen data set. Specifically for this study, five-unit space data were selected and discarded from the raw train data. As such, the total number of signal data is about five less than the total number of the raw train data set.

Prior to verifying the effectiveness of the proposed method in feature selection optimization, the basic T-method prediction model involving all features was developed using the normalized signal data set. The model parameters required in formulating the prediction model, such as the proportional coefficient, β and SNR, η , were computed for each dataset. Table 2 shows the example of proportional coefficient, β and SNR, η parameters computed for the cooling load dataset. Any feature with a negative value of SNR will be replaced with a zero value, as shown in feature X6 in Table 2. This is in accordance with the condition given in (4). Next, the optimal parameter setting for the NIBBA is determined for each dataset. Specifically, for this experimental study, the Taguchi method was employed by utilizing an L_{27} orthogonal array in designing the experiment involving six parameters with three levels, as shown in Table 3. Each parameter's level was obtained from various studies of past research employing the Binary Bat algorithm. For each combination in L_{27} array, the experiment run repeated three times, and the average value of integrated estimate SNR (db) was used as the response to determine each feature's significant level except for the Abalone dataset. Since every combination in the experimental run converged to an optimal objective function which is the maximum value of integrated estimate SNR, the response for the Abalone dataset is set as convergence rate computed using (27). With the utilization of Minitab software, the parameter's level with the highest signal-to-noise ratio was selected as the optimal parameter setting. The optimal parameter setting for each dataset is as shown in Table 4. Next, in obtaining the optimal input feature using the proposed T-method with the NIBBA approach utilizing the optimal parameters setting, 20 independent runs were executed with 500 internal iterations for each run. The objective function of the proposed T-method with the NIBBA algorithm is to maximize the integrated estimate SNR value, which returns the optimal combination of input features. The selection of the final optimal feature combination is based on the 50% and more feature appearance in the optimal combination of every run, as practiced in [14].

Table 1. Benchmark datasets

Dataset	No. of sample	No. of feature	Train data set		Validation data set (30%)
			70%	Signal data	
Abalone	4177	7	2924	2919	1253
Concrete	1030	8	721	716	309
Cooling load	768	8	538	533	230
Body fat	252	14	177	172	75

Table 2. Computed model parameters for cooling load dataset

Parameter	X1	X2	X3	X4	X5	X6	X7	X8
β	0.007	-6.342	1.772	-4.057	0.166	0.001	0.003	0.012
η	0.005	0.005	0.002	0.011	0.007	-1.897	0.0001	2.404
Corrected η	0.005	0.005	0.002	0.011	0.007	0	0.0001	2.404

Table 3. Experimental parameters settings

Parameter	Level 1	Level 2	Level 3
Population size	10	25	30
Frequency	[0, 1]	[0, 2]	[0.8, 1]
Loudness	0.25	0.5	0.75
Pulse rate	0.01	0.5	0.9
Alpha, α	0.1	0.7	0.9
Gamma, γ	0.6	0.9	0.95

Table 4. NIBBA parameters setting

Parameter	Abalone	Concrete	Cooling Load	Body Fat
Population size	25	10	10	25
Min Freq, F_{min}	0	0	0	0
Max Freq, F_{max}	2	2	2	1
Loudness, A_0	0.5	0.25	0.25	0.75
Pulse rate, r_0	0.5	0.9	0.5	0.01
Alpha, α	0.9	0.1	0.1	0.9
Gamma, γ	0.95	0.6	0.6	0.6

$$\text{Simple convergence rate, } C_r = \frac{(\text{Iteration}_{Max} - \text{Iteration}_{start\ converge})}{\text{Iteration}_{Max}} \times 100\% \quad (27)$$

Upon obtaining the optimal subset of input features, a feature reduction rate is computed using (28) to evaluate the reduction ability of feature selection [27]. Subsequently, the prediction accuracy is computed using the mean absolute error (MAE) error metric as the performance measure in quantifying the model's accuracy using the validation data set, as shown in (29).

$$\text{Reduction rate, } R_r = \frac{\# \text{ original features} - \# \text{ selected features}}{\# \text{ original features}} \times 100\% \quad (28)$$

$$\text{Mean Absolute Error (MAE)} = \frac{1}{l} \sum_{i=1}^l |M_i - \widehat{M}_i| \quad (29)$$

where M_i is the actual output of the validation data set, \widehat{M}_i is the predicted output value for the validation data set, and l is the number of validation data set. A comparison study is then conducted to compare the prediction performance between the T-method with full feature, T-method with OA, and T-method with NIBBA. Finally, hypothesis testing using a paired t-test statistical methodology is performed to determine whether there are significant differences between the mean outcome of each approach, as formulated in (30). A p -value is established to decide whether to accept or reject the null hypothesis at a significance level of 0.05 (5%) [28]. Throughout the study, a MATLAB R2020a programming application software was utilized in constructing and executing the algorithms on a laptop-type computer powered by an Intel Core i5-8250U central processing unit, 4 Gigabytes of random-access memory, and 1 Terabyte of storage capacity.

$$\begin{aligned} H_0: \mu_1 &= \mu_2; \text{ mean difference is equal (no difference)} \\ H_1: \mu_1 &\neq \mu_2; \text{ mean difference is not equal (difference)} \end{aligned} \quad (30)$$

3.2. Experimental results

Table 5 shows the optimal number of features with their respective combination (in the parenthesis) and reduction rate, R_r for all three approaches obtained using the train data set. Apparently, for the T-method with full features, all original features are used in formulating the prediction model. On the contrary, the T-method with OA and T-method with NIBBA approaches recorded a reduction in the total number of input features with different optimal feature combinations. In general, a reduction in the total number of features indicates that a less complex T-model prediction model was achieved through the T-method with OA and T-method with NIBBA, which potentially improved the prediction accuracy and fastened the computation time. It can also be concluded that both the T-method with OA and T-method with NIBBA approaches successfully identify insignificant features and offer only a subset of significant features to be incorporated in the prediction model based on their respective methodology.

Table 6 shows the prediction accuracy in terms of MAE value for the three approaches obtained using optimal features combination on the validation data set. The percentage value inside the parenthesis is the percent enhancement of the T-method with OA and T-method with NIBBA against the T-method with full features. Obviously, the proposed T-method with NIBBA recorded the best MAE value for all case studies, indicating the optimality of the prediction model achieved when only significant features were incorporated in the model. Nevertheless, the conventional approach also recorded MAE enhancement on the prediction accuracy, suggesting that a feature selection process is important in a multivariate dataset. From the perspective of the ability in obtaining the model's optimality, when more combination is generated and analyzed as in the NIBBA, the probability of obtaining an optimal feature subset that contributes to the enhancement of the T-method prediction accuracy is greater as opposed to conventional fixed and limited combination to be assessed.

Table 7 shows the result of paired t-test hypothesis testing to determine whether there are significant differences in the mean of predicted outcome between the T-method with OA and T-method with NIBBA

approaches using the validation data set. The recorded p -value for the concrete, cooling load, and body fat datasets was less than the significance level of 0.05, resulting in the rejection of the null hypothesis (H_0). Thus, it can be concluded that there were significant differences between the outcome of both approaches. However, for the Abalone dataset, the p -value of more than 0.05 failed to reject the null hypothesis and suggested the differences between both approaches are not statistically significant.

Table 5. Optimal features combination

Dataset	Item	T-method + full features	T-method + OA	T-method + NIBBA
Abalone	No. of feature	7	3	2
	Optimal combination	all	(2, 3, 7)	(3, 7)
	Reduction rate, R_r	-	57.1%	71.4%
Concrete	No. of feature	8	6	6
	Optimal combination	all	(1, 3, 4, 5, 7, 8)	(1, 2, 4, 6, 7, 8)
	Reduction rate, R_r	-	25.0%	25.0%
Cooling load	No. of feature	8	4	4
	Optimal combination	all	(3, 4, 7, 8)	(1, 3, 7, 8)
	Reduction rate, R_r	-	50.0%	50.0%
Body fat	No. of feature	14	4	3
	Optimal combination	all	(1, 2, 4, 7)	(1, 2, 7)
	Reduction rate, R_r	-	71.4%	78.6%

Table 6. Prediction accuracy (MAE) on the validation data set

Dataset	T-method + full features	T-method + OA	T-method + NIBBA
Abalone	3.65	3.23 (11.5%)	3.16 (13.4%)
Concrete	16.36	16.15 (1.0%)	15.19 (7.2%)
Cooling load	8.08	5.95 (26.4%)	5.40 (33.2%)
Body fat	0.40	0.25 (37.5%)	0.18 (55.0%)

Table 7. T-test result

Dataset	T-value	p -value
Abalone	1.89	0.059
Concrete	-5.57	0.000
Cooling load	-5.24	0.000
Body fat	-4.29	0.000

4. CONCLUSION

This paper proposed the integration of the nearest integer-based binary bat algorithm with Taguchi's T-method as the feature selection optimization. The results from the experimental study using benchmark datasets show that the integration is feasible, and the NIBBA algorithm is capable of searching for the optimal feature subset affecting the T-method prediction accuracy. It was observed that the process of feature selection, either through conventional OA or NIBBA algorithm, successfully discarded insignificant features, resulting in a less complex model. Moreover, the T-method with NIBBA successfully obtained a better subset of optimal features as compared to the conventional T-method with OA based on MAE results recorded using the validation data set. In addition, the results from paired t-test hypothesis testing indicate that the difference between the T-method with OA and T-method with NIBBA outcome is statistically significant and not a chance of coincidence, except for the Abalone case study. In general observation, it was found that the recursive and adaptive feature embedded in the binary bat algorithm drives the search for an optimal solution. As for the nearest integer discretization approach, it allows the bat to globally explore the continuous-valued solution space and exploit the continuous-valued best solution during local search since the discretization is executed at the destination or bat's position. Furthermore, NIBBA operates at a low computational cost. In conclusion, the integration between T-method and NIBBA algorithm is beneficial by offering better T-method prediction accuracy.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Teknologi Malaysia and the Ministry of Higher Education, Malaysia for the Fundamental Research Grant Scheme (FRGS/1/2021/TK0/UTM/02/45) that have supported this research




REFERENCES

- [1] V. Kumar and M. L., "Predictive analytics: a review of trends and techniques," *International Journal of Computer Applications*, vol. 182, no. 1, pp. 31–37, Jul. 2018, doi: 10.5120/ijca2018917434.
- [2] M. Asiah, K. Nik Zulkarnaen, D. Safaai, M. Y. Nik Nurul Hafzan, M. Mohd Saberi, and S. Siti Syuhaida, "A review on predictive modeling technique for student academic performance monitoring," *MATEC Web of Conferences*, vol. 255, p. 03004, Jan. 2019, doi: 10.1051/mateconf/201925503004.
- [3] A. K. Waljee, P. D. R. Higgins, and A. G. Singal, "A primer on predictive models," *Clinical and Translational Gastroenterology*,




- vol. 5, no. 1, p. e44, Jan. 2014, doi: 10.1038/ctg.2013.19.
- [4] S. Teshima and Y. Hasegawa, *Quality recognition & prediction: smarter pattern technology with the Mahalanobis-Taguchi system*. New York: Momentum Press, 2012.
 - [5] V. L. Chetana, S. S. Kolisetty, and K. Amogh, "A short survey of dimensionality reduction techniques," in *Recent Advances in Computer Based Systems, Processes and Applications*, CRC Press, 2020, pp. 3–14.
 - [6] D. M. Hawkins, "Discussion," *Technometrics*, vol. 45, no. 1, pp. 25–29, Feb. 2003, doi: 10.1198/004017002188618653.
 - [7] B. Abraham and A. M. Variyath, "Discussion," *Technometrics*, vol. 45, no. 1, pp. 22–24, Feb. 2003, doi: 10.1198/004017002188618644.
 - [8] W. H. Woodall, R. Koudelik, K.-L. Tsui, S. B. Kim, Z. G. Stoumbos, and C. P. Carvounis, "A review and analysis of the Mahalanobis-Taguchi system," *Technometrics*, vol. 45, no. 1, pp. 1–15, Feb. 2003, doi: 10.1198/004017002188618626.
 - [9] S. B. Kim, K. L. Tsui, T. Sukchotrat, and V. C. P. Chen, "A comparison study and discussion of the Mahalanobis-Taguchi system," *International Journal of Industrial and Systems Engineering*, vol. 4, no. 6, p. 631, 2009, doi: 10.1504/IJISE.2009.026768.
 - [10] J. Nourmohammadi-Khiarak, M.-R. Feizi-Derakhshi, F. Razeghi, S. Mazaheri, Y. Zamani-Harghalani, and R. Moosavi-Tayebi, "New hybrid method for feature selection and classification using meta-heuristic algorithm in credit risk assessment," *Iran Journal of Computer Science*, vol. 3, no. 1, pp. 1–11, Mar. 2020, doi: 10.1007/s42044-019-00038-x.
 - [11] H. Yadav, A. C. Kumari, and R. Chhikara, "Feature selection optimisation of software product line using metaheuristic techniques," *International Journal of Embedded Systems*, vol. 13, no. 1, p. 50, 2020, doi: 10.1504/IJES.2020.108284.
 - [12] P. Agrawal, H. F. Abutarboush, T. Ganesh, and A. W. Mohamed, "Metaheuristic algorithms on feature selection: a survey of one decade of research (2009-2019)," *IEEE Access*, vol. 9, pp. 26766–26791, 2021, doi: 10.1109/ACCESS.2021.3056407.
 - [13] N. Harudin, J. K. R. M. Nabil Muhtazaruddin, R. F. W. Zuki Azman Wan Muhammad, and N. Jaafar, "Artificial bee colony for features selection optimization in increasing T-method accuracy," *International Journal of Engineering & Technology*, vol. 7, no. 4.35, p. 885, Nov. 2018, doi: 10.14419/ijet.v7i4.35.26276.
 - [14] N. Harudin *et al.*, "Binary bitwise artificial bee colony as feature selection optimization approach within Taguchi's T-method," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–10, May 2021, doi: 10.1155/2021/5592132.
 - [15] Z. M. Marlan, K. R. Jamaludin, F. Ramlie, and N. Harudin, "Determination of optimal unit space data for Taguchi's T-method based on homogeneity of output," *Open International Journal of Informatics (OIJI)*, vol. 7, pp. 167–179, 2019.
 - [16] G. Taguchi, S. Chowdhury, and Y. Wu, *Taguchi's Quality Engineering Handbook*. Livonia, Michigan: John Wiley & Sons, Inc., 2004.
 - [17] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Studies in Computational Intelligence*, vol. 284, 2010, pp. 65–74.
 - [18] X. S. Yang, *Nature-inspired metaheuristic algorithm*, Second Edi. United Kingdom: Luniver Press, 2010.
 - [19] R. Y. M. Nakamura, L. A. M. Pereira, D. Rodrigues, K. A. P. Costa, J. P. Papa, and X.-S. Yang, "Binary bat algorithm for feature selection," in *Swarm Intelligence and Bio-Inspired Computation*, Elsevier, 2013, pp. 225–237.
 - [20] S. Mirjalili, S. M. Mirjalili, and X.-S. Yang, "Binary bat algorithm," *Neural Computing and Applications*, vol. 25, no. 3–4, pp. 663–681, Sep. 2014, doi: 10.1007/s00521-013-1525-5.
 - [21] B. Crawford, R. Soto, G. Astorga, J. Garcia, C. Castro, and F. Paredes, "Putting continuous metaheuristics to work in binary search spaces," *Complexity*, vol. 2017, pp. 1–19, 2017, doi: 10.1155/2017/8404231.
 - [22] H. Yoshida, K. Kawata, Y. Fukuyama, S. Takayama, and Y. Nakanishi, "A particle swarm optimization for reactive power and voltage control considering voltage security assessment," *IEEE Transactions on Power and Energy*, vol. 119, no. 12, pp. 1462–1469, 1999, doi: 10.1541/ieejpes1990.119.12_1462.
 - [23] Z. A. E. M. Dahi, C. Mezioud, and A. Draa, "Binary bat algorithm: on the efficiency of mapping functions when handling binary problems using continuous-variable-based metaheuristics," in *5th International Conference on Computer Science and Its Applications (CIIA)*, Saida, Algeria, 2015, pp. 3–14.
 - [24] M. Sevkli and A. R. Guner, "A continuous particle swarm optimization algorithm for uncapacitated facility location problem," in *Communications in Computer and Information Science*, vol. 147 CCIS, Berlin Heidelberg: Springer-Verlag, 2006, pp. 316–323.
 - [25] A. Frank and A. Asuncion, "UCI machine learning repository," Irvine, CA: University of California, School of Information and Computer Science., 2010. <http://archive.ics.uci.edu/ml>.
 - [26] R. W. Johnson, "Carnegie Mellon University," in *Green Education: An A-to-Z Guide*, 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2011.
 - [27] A. A. A. Al-Hajjana, "Text feature selection using enhanced binary bat algorithm," Universiti Kebangsaan Malaysia, 2018.
 - [28] J.-B. du Prel, G. Hommel, B. Röhrig, and M. Blettner, "Confidence interval or P-value? Part 4 of a series on evaluation of scientific publications," *Deutsches Ärzteblatt international*, vol. 106, no. 19, pp. 335–339, May 2009, doi: 10.3238/arztebl.2009.0335.

BIOGRAPHIES OF AUTHORS






Zulkifli Marlah Marlan    received a Bachelor degree in Mechanical (Industrial) Engineering from Universiti Teknologi Malaysia in 2008 and a Master degree in Engineering Management from Universiti Putra Malaysia in 2014. He is a Graduate Member of Board of Engineers, Malaysia. Currently, he is conducting research as a Phd postgraduate student in Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Malaysia on Taguchi's T-method for prediction. His research areas include Mahalanobis-Taguchi system, Metaheuristic algorithm and robust statistics. He can be contacted at email: zulkifli49@graduate.utm.my.






Khairur Rijal Jamaludin    is an Associate Professor at Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Malaysia. He received a Bachelor degree in Mechanical Engineering from Universiti Teknologi Malaysia in 1996, a Master of Science degree in Advanced Mechanical Engineering from University of Warwick in 1998 and a Ph.D. from Universiti Kebangsaan Malaysia in 2009. He is a Professional Technologist (Ts.) of Malaysia Technologies Board and a member of various professional membership such as Graduate Member Board of Engineers, Malaysia, Graduate Member Institution of Engineers Malaysia, Member of American Powder Metal Institute and Quality Engineering Society, Japan. His research interests are primarily in the area of robust engineering (Taguchi Method and Mahalanobis-Taguchi system), remanufacturing, powder metallurgy and particulate materials, quality engineering and quality management, Lean / Toyota Production System, and end of life vehicles/ vehicles recycling, where he is the author/co-author of over 50 research publications. He can be contacted at email: khairur.kl@utm.my.



Faizir Ramlie    is a senior lecturer at Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Malaysia since 2017. He received a Bachelor degree in Mechanical Engineering from University of Queensland Australia in 1997, a Master degree in Mechanical Engineering Specializing in Advanced Manufacturing Technology and a Ph.D. from Universiti Teknologi Malaysia in 2016. Experienced engineer over 15 years of experience in various multinational companies in Malaysia, involving product development and manufacturing. His research interests are primarily in the area of pattern recognition using Taguchi Method and artificial intelligence, Mahalanobis-Taguchi system, robust engineering and manufacturing technology. He can be contacted at email: faizir.kl@utm.my.



Nolia Harudin    is a senior lecturer at Univeristi Tenaga Nasional, Malaysia since 2013. She received a Bachelor degree in Mechanical (Industrial) Engineering from Universiti Teknologi Malaysia in 2008, a Master degree in Mechanical Engineering from Universiti Teknologi Malaysia in 2012 and a Ph.D. from Universiti Teknologi Malaysia in 2020. She is a Graduate Member of Board of Engineers, Malaysia. Her research interests are Taguchi Method, Mahalanobis-Taguchi system, pattern recognition, energy efficiency, robust engineering and quality control. She can be contacted at email: nolia@uniten.edu.my.