Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Multiple Infill Sampling Strategy using Multi-Surrogate Modelling for Global Optimization Problem

Che Munira Che Razali[1,2,*], Shahrum Shah Abdullah[2], Amrul Faruq[3]

[1] Department of Electrical Engineering, Politeknik Ungku Omar, 31400 Ipoh, Perak, Malaysia
[2] Department of Electronic System Engineering, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia
[3] Universitas Muhammadiyah Malang, Kota Malang, Jawa Timur 65113, Indonesia

ABSTRACT

In recent years, optimization using a surrogate model or metamodel received great scholarly attention in solving computer simulation problems. Surrogate models are fast approximation, high-fidelity models and better accuracy in the prediction of the model. The infill sampling strategy is the one-way method to refine the surrogate model and improve the accuracy of the model. This paper proposed a multiple adaptive sampling strategy using two (2) surrogate models, Radial Basis Function and Kriging. The proposed method of the multi-surrogate model predicts two sample points with a combination of DBSCAN clustering as the initial processing of training points. This approach helps to improve the performance of the algorithm in terms of accuracy by calculating root mean square error (RMSE). The contribution of an algorithm proposed 2 sample points prediction at one iteration instead of previous research only predicting one sample for each iteration. The algorithm was tested and demonstrated using low dimension test function benchmark from previous work.

Keywords:
Multiple infill sampling; Multi-surrogate model; Input sample prediction

1. Introduction

Computational modelling plays a pivotal role in scientific and technological domains as it involves simulating complex real-world problems that demand computationally intensive algorithms. Surrogate models, which are simplified functional approximations of intricate models, present a valuable approach to facilitating engineering analysis of complex systems by substantially reducing computational costs. Addressing challenges in computational simulation, integrating a surrogate model utilizing the Design of Experiment (DOE) and Response Surface Method (RSM) has proven to be a powerful and efficient tool. This surrogate model involves the creation of a compact analytical model that effectively approximates the intricate analysis, providing a practical and resource-efficient solution. This model can serve as a "surrogate" or close substitute for the high-precision analysis while ensuring accuracy. Surrogate modelling techniques like Kriging, radial basis function

* *Corresponding author.*
E-mail address: chemunira@puo.edu.my

https://doi.org/10.37934/araset.62.3.102117

(RBF), polynomial response surface (PRS), support vector regression (SVR) and multivariate adaptive spline (MARS) are commonly used [1,2]. Surrogate modelling, also known as metamodeling, is effectively used to obtain a function *f(x*) that approximates the function *f*. The optimization process to identify the best feasible value *f(x),* in a predetermined period is frequently defined by the number of available queries to the black box. The advantage of using a metamodel is reducing the computational cost necessary to approximate the numerical model output [3]. In another field of research, the surrogate model is used to develop a framework for detecting SQL Injection using machine learning and classification methods such as Random Forest Classifier, Gradient Boosting Classifier, SVM and ANN [4]. Recently, multiple surrogates have been widely used to replace the expensive computational model in design and real-world optimization problems.

Despite the aggressive development of algorithms using surrogate models for prediction sample points, previous researchers also propose a new method combining several surrogate models to enhance the performance and for robustness to model assumption. Wang *et al.,* [5] propose a parallel infill sampling criterion for EGO. Instead of adding a single sample point in each updating cycle as the original EGO does, the EGO with the proposed method can obtain an arbitrary number of new sample points per cycle, which will be evaluated in parallel. The results show that the optimization efficiency is significantly improved compared to the serial EGO and the effectiveness is promoted in contrast to the existing parallel infill criteria. Another research paper for multiple infill sampling method proposed by Chao *et al.,* [6] for low-fidelity model and the multi-infill strategy are utilized in this approach. Low-fidelity data is employed to provide a good global trend for model prediction and multiple sample points chosen by different infill criteria in each updating cycle are used to enhance the exploitation and exploration ability of the optimization approach. Take the advantages of low- fidelity model and the multi-infill strategy and no initial sample for the high-fidelity model is needed. The result shows that more than 60% of the computational cost is saved compared with ordinary Kriging using the same infill strategy. Multiple infill sampling aims to improve accuracy and computational time instead of predicting one sample at each iteration.

A new strategy for infill sampling called the multi-point infill sampling strategy has been introduced by Aburashed *et al.,* [4]. This strategy locates new promising points near optimal points to speed up the optimization process, where a hybrid and adaptive promising sampling (HAPS) method and a multi-start sequential quadratic programming (MSSQP) method are used alternately. The proposed method, the multi-surrogates and multi-points infill strategy-based global optimization (MSMPIGO), has been tested on eighteen unconstrained optimization problems, six nonlinear constrained engineering problems and one air foil design optimization problem. Strong evidence of a multi-point strategy is also supported by Song *et al.,* [7], which proposes a Kriging-based global optimization using a multi-point infill sampling criterion. This method uses an infill sampling criterion which obtains multiple new design points to update the Kriging model by solving the constructed multi-objective optimization problem in each iteration. A simulation-based optimization based on the 445 bus lines in Beijing City is employed to test the performance of the proposed algorithm. However, the method in this paper is the lack of more practical applications. In the future, the author will focus more on black-box optimization combined with time-consuming simulations of real-world traffic problems and continuously improve the algorithm in practical applications. A new strategy for infill sampling called the multi-point infill sampling strategy has been introduced by Aburashed *et al.,* [4]. This strategy locates new promising points near optimal points to speed up the optimization process, where a hybrid and adaptive promising sampling (HAPS) method and a multi-start sequential quadratic programming (MSSQP) method are used alternately. The proposed method, the multi-surrogates and multi-points infill strategy-based global optimization (MSMPIGO), has been tested on eighteen unconstrained optimization problems, six nonlinear constrained engineering problems and

one air foil design optimization problem. In 2018, Song *et al.,* [7] proposed multiple-update-infill sampling for Kriging using a minimum energy design to improve the global quality of the surrogate model. The method was evaluated with other multiple-update-infill sampling methods in terms of convergence, accuracy, sampling efficiency and computational cost. During the development of the algorithm, it is vital to remember that multiple infill sampling with multi-surrogate also has some drawbacks, such as increased computational time and the risk of overfitting.

There is consensus among scientists about the drawbacks and strengths of each surrogate model. The algorithm development is based on optimization problems and there's no free lunch theorem for algorithm solving the problems or case studies. RBF for infill sampling can result in a smooth, continuous function approximating the modelled underlying function. This function can be used to make predictions at points where samples have not been taken and can help to identify regions of the parameter space where the optimization should be focused. On the other hand, infill sampling using Kriging can provide more information about the underlying function, including a measure of the uncertainty in the model. Combining Kriging and Radial Basis Functions for a hybrid approach can improve the algorithm's performance and accuracy. Hwang *et al.,* [8] conducted research using surrogate-assisted global and local searches with assisted hybrid evolutionary optimization performed in sequence at each generation to balance the exploration and exploitation is efficient for solving the low- and medium-dimensional expensive optimization problems compared to the other six state-of-the-art surrogate-assisted evolutionary algorithms. Yu *et al.,* [10] suggested a brand-new model management strategy based on multi-RBF parallel modelling technology in this paper. The proposed approach aims to adaptively select a high-fidelity surrogate from a pre-specified set of RBF modelling techniques during the optimization process. At each evolutionary interaction, the most promising RBF surrogate was employed to help the neighbourhood field optimizer (NFO) perform fitness evaluation and the proposed algorithm is named aRBF-NFO. A hybrid model for the surrogate model is not only proposed for SM only but also in combination with the metaheuristic method showing good performance and overcoming the drawback of the method.

Multiple infill sampling is one of the popular and well-developed techniques to handle the issue of computational time and accuracy of the metamodel. However, other techniques can also be implemented similarly to multiple infill strategies, namely the batch infill technique. Habib *et al.,* [11] have developed a new method for sampling multiple locations during each iteration. They have proposed a multi-objective (MO) formulation to maximize the expected improvement and the distance from previously evaluated solutions. Another research paper by the same author also presents a multi-objective formulation to deal with such classes of problems, wherein instead of a single solution, a batch of solutions is identified for concurrent evaluation. The strategies use different objectives depending on the archive of the evaluated solutions [12]. Researchers in computational statistics used the R package software to develop criteria for batch-sequential inversion, which allows advanced users to distribute function evaluations across clusters or clouds of machines in parallel. This software package uses the KrigInv present tutorial to make it easy for people unfamiliar with kriging to use the box and clarify the strengths and weaknesses of these metamodel-based inversion methods [13]. According to studies discussed by Habib *et al.,* [11] in this section, batch infill sampling was implemented for multi-objective optimization problems. However, in previous research papers, other authors used the term multiple infill sampling when solving multi-objective problems.

Xing *et al.,* [14] demonstrate Kriging with parallel computing to improve computer efficiency and solve global solutions. His work proposes a global optimization strategy based on the Kriging surrogate model and parallel computing depending on the multipeak characteristics of the expected improvement (EI) function. Compared with the conventional EI criterion and the parallel constant

Liar criterion, the proposed PEI-R method considerably improves the optimization efficiency and solution accuracy. At the same time, Chen *et al.,* [15] compares the common efficient parallel infill sampling criterion. In addition, the pseudo-expected improvement (EI) criterion is introduced to minimize the predicted (MP) criterion and the probability of improvement (PI) criterion, which helps to improve the problem of the MP criterion that is easy to fall into local optimum. An adaptive distance function is proposed, which is used to avoid the concentration problem of update points and improves the global search ability of the infill sampling criterion. Another work by Yang *et al.,* [16] proposes five alternatives of Probability of Improvement (PoI) with multiple points in a batch (q-PoI) for multi-objective Bayesian global optimization (MOBGO), taking the covariance among multiple points into account. Efficient global optimization (EGO) is another name for kriging metamodel with Expected Improvement infill strategy. Based on recent studies in this section, multiple infill sampling, batch infill methods or parallel infill sampling methods are recent approaches for surrogate model updating points to solve optimization problems. In a comprehensive literature review of the surrogate model, Hafka *et al.,* [17] focused on in this review is how different algorithms balance exploration and exploitation. This author agreed that methods that provide easy parallelization, like multiple parallel runs or methods that rely on a population of designs for diversity, deserve more attention based on his review. Based on the strength and widespread research on surrogate models with various infill sampling techniques, this paper proposes a method with a multi-surrogate model with multiple infill techniques combined with DBSCAN (Density-based spatial clustering of applications with noise) as a clustering method at the pre-processing stage. The algorithm's performance is measured and determined by evaluating the root mean square error (RMSE) value for the previous benchmark mathematical test function.

2. Methodology of Multiple Adaptive Sampling Multi-Surrogate Model

The main objective of infill sampling criteria (ISC) is to extract information from surrogate models to identify potentially interesting areas for model refinement (and possibly feasibility), striking a balance between model exploitation and exploration. Consequently, the goal of an infill search criterion is to extract the maximum amount of information from the fewest number of samples by striking a balance between sample size and the amount of data to be extracted between:

i. Exploiting regions of the design space where the surrogate model indicates there might be a minimizer.
ii. Exploring under-sampled areas with high estimated surrogates' error.
iii. Searching for feasible regions, i.e., Regions where all constraints are satisfied.

Conventional one-shot sampling uses the design of experiment (DoE) method. The DoE method effectively optimizes the number of experiments and parameter experiments. However, DoE was typically used in computer experiments to generate the initial sampling point.

Metamodeling is a computational optimization technique which involves four (4) stages:

i. sampling technique
ii. approximation function
iii. obtaining a new sample
iv. refining the metamodel.

Lin *et al.,* [18] summarize each metamodel's use and fitting alternatives. For low-dimensional problems, the Response Surface Method (RSM) is well-established, user-friendly and suitable. On the other hand, the Neural Network metamodel excels in addressing serious nonlinear problems requiring substantial samples, making it an excellent choice for deterministic applications. The Kriging metamodel stands out as a flexible and highly effective option for low-dimensional scenarios. In the subsequent section, we will delve into the proposed method of infill sampling, employing two surrogate models to enhance the optimization process further.

2.1 DBSAN Classifier

DBSCAN is an algorithm for density-based clustering initially designed to discover clusters, *C*, of arbitrary shapes in spatial data[19]. The algorithm takes two input parameters:

i. the radius of a hypersphere drawn around each point, known as a neighbourhood,
ii. MinPts, the minimum number of points in the neighbourhood, must be defined as a part of a cluster, including the current point.

The DBSCAN algorithm detects clusters with a minimum data density specified by the user for initial outlier filtering. The k-means algorithm divides these clusters into *k* groups with uniform distribution. The group centroids are then utilized for data reduction by replacing each group with the actual data point closest to the identified centroid [20]. This paper uses DBSCAN to cluster training data twofold:

i. global search, which is the entire region far from the optimal point
ii. local search, which is the sample located near the optimal point. After clustering the samples, kriging metamodel performs the local search while RBF metamodel performs global search of training data.

2.2 RBF with Maximin Distance for Global Search Infill Sample
2.2.1 Radial basis function

The Radial Basis Function Neural Network (RBFNN) stands out as a robust algorithm in Artificial Neural Networks (ANN). It comprises three feed-forward, fully connected layers, employing RBFNN as the exclusive nonlinearity in the hidden layer neurons. Unlike the hidden layer, the output layer of RBFNN lacks nonlinearity and utilizes solely weighted connections. Furthermore, the connections from the input to the hidden layer remain unweighted. RBFNN boasts superior approximation capabilities, featuring a simpler network architecture and a faster learning algorithm [21]. Figure 1 shows the architecture of RBFNN which consists of input, hidden layer and output.

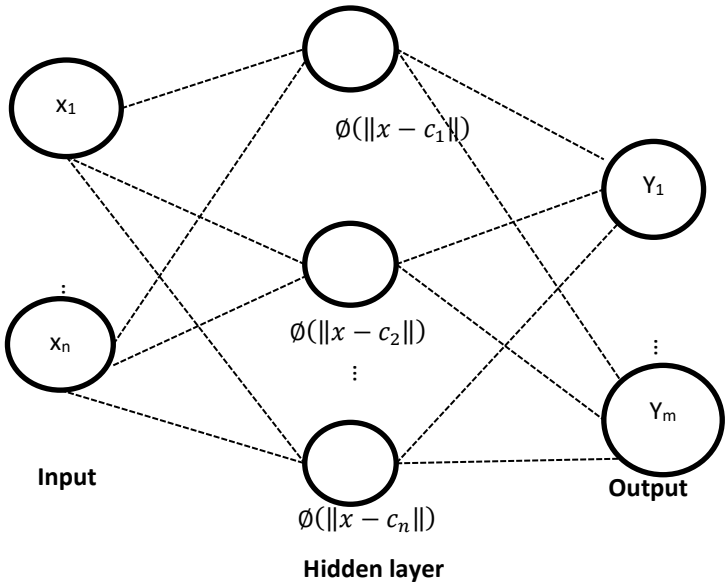

Fig. 1. RBF architecture

The RBFNN model is then expressed as a linear combination of the basis function across all the training points *m* as given by the equation. In Eq. (1), w_i are the generic weights of the basis function. The weight is evaluated by training points xi and corresponding function values $f(x_i)$. The matrix of the basis function values at the training points is described in Eq. (2).

$$f(x) = \sum_{i=1}^{N} w_i \emptyset\,(x) \tag{1}$$

Denotes that w_i is weight while hi(x) is a set of K arbitrary nonlinear functions known as radial basis function and ‖. ‖denotes a norm of Euclidean distance.

$$\emptyset = \begin{bmatrix} \emptyset_{11} & \emptyset_{12} & \dots & \emptyset_{1N} \\ \emptyset_{11} & \emptyset_{22} & \cdots & \emptyset_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \emptyset_{k1} & \emptyset_{k2} & \dots & \emptyset_{NN} \end{bmatrix} \tag{2}$$

A typical radial basis function is the Gaussian expressed by the equation below:

$$\emptyset_i\,(x) = e^{\left(-\frac{(x-c)}{\beta^2}\right)} \tag{3}$$

where *X* is the input, *C* is the centre and *β* is the spread parameter. Table 1 list of equation activation function that can be implemented for RBFNN. The activation function can influence the output prediction of the algorithm.

All the input data are represented by the input matrix *X* and the output data are represented by the output vector *y* as below:

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_N^1 \\ x_1^2 & x_2^2 & \cdots & x_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \dots & x_N^M \end{bmatrix} \tag{4}$$

The optimum value in the second layer weight can be found using the least square formula below:

$$\widehat{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = (\emptyset^T \emptyset)^{-1} \emptyset^T \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \tag{5}$$

Table 1
Activation function for radial basis function

Equation	Activation Function
$\varphi(r,\sigma) = e^{\frac{-r^2}{\sigma^2}}$	Gaussian
$\varphi(r) = r^3$	Cubic
$\varphi(r) = r^2 \log(r)$	Thin Plate spline
$\varphi(r) = \frac{1}{r+1}$	Cauchy
$\varphi(r,\beta) = \sqrt[2]{r^2 + \beta}$	Multiquadratic
$\varphi(r,\beta) = \frac{1}{\sqrt[2]{r^2 + \beta}}$	Inverse multi-quadratic
$\varphi(r) = r$	Linear

The approximation proposed using polynomial regression and RBF neural network to build a rocket aerodynamic discipline surrogate model are both valid, while the surrogate model adopting the RBF neural network gets better results and profits from its adaptability for more types of data [22]. RBF consists of two (2) important parameters that determine the output function approximation. Selection of the suitable parameter of centre and spread gives a good predicted output value. The spread value in this problem is fixed as constant and selected arbitrarily based on the minimum error criteria. K-means clustering to find centres for basis function in a fashion that reflects the distribution of input vectors over the input space [18]. Some literature that focuses on the density of each point provided insight for our work to progress in using distance-based weight for selecting better centres for RBFN training. In this paper, Fuzzy c-mean (FCM) is a clustering method used to determine the optimal centre, while the spread parameter of the RBF model is calculated by using the nearest neighbour based on percentage distance from the centre.

2.2.2 Maximin distance approach

The maximin distance criterion was proposed by Johnson in 1990 for computer experiments. Given the existing sample set *Xp*, the maximin distance approach is to select a new sample *Xc* to maximize the minimum distance between any two-sample point in the sample set $X_A = Xc$, $X_A = X_c \cup X_P$. The equation for this method is as below:

$$\max Xc \left[\min_{1 \le i \ge m, 1 \le j \ge 1+m} \left(d\left(X_{Ci}, X_{Aj}\right) \right) \right] \tag{6}$$

Max–min distance designs tend to cover the design space as much as possible because no two points should be too close to each other. The distance min $x \in C$ $d(x, C \backslash x)$ can be interpreted to indicate the degree of covering by the candidate design *C*. On the other hand, min-max distance designs tend to spread out in the design space as uniformly as possible because we can interpret that points outside of *C* pull out C as much as possible. These distance-based criteria yield a uniform design and fill the design space as much as possible.

2.3 Kriging with Expected Improvement Optimized by Grey Wolf Algorithm for Local Search
2.3.1 Kriging metamodel

For design point $X = [x_1, \dots, x_m]^N$, where $X \in R^{mxn}$ and response $Y = [Y_1, \dots, Y_m]^N$ with $Y \in R^{mxn}$, the kriging model is the combination of the trend term and the deviation term:

$$\hat{y}(x) = f(x) + Z(x) \tag{7}$$

where $\hat{y}(x)$ is the objective estimation of the Kriging model, *f*(*x*) is the known function of *x*, which is similar to the response surface polynomial model, provides a global optimization model in design space *z*(*x*) is a random process, the covariance can be expressed as formula below:

$$cov\left[Z(x)^i, Z(x)^j\right] = \sigma R\left[R(x^i, x^j\right] \tag{8}$$

R is the correlation matrix and $R(X^i$ and $X^j)$ is the correlation function of any two-sample points X^i and X^j. There are various correlation functions, such as the exponential, Gaussian and spline functions for kriging metamodel.

In summary, using Kriging metamodels, support vector machines and the expected improvement criterion can enhance the efficiency and accuracy of metamodel-based design optimization algorithms. These techniques have been applied in various fields, including topology optimization, aerodynamic shape optimization and robust design optimization.

2.3.2 Expected improvement

The EI criterion is developed by assuming that the uncertainty in the predicted value, $\hat{y}(x)$ at a position x, can be described as a normally distributed random variable Y(x). The Kriging interpolator $\hat{y}(x)$, is assumed to be the mean of this random variable while the variance is considered to be given by the Kriging mean square error, $S^2(x)$. The improvement of the unsampled point beyond the current best-observed value, y_{min}, is also a random value, which can be expressed as:

$$I(x) = max(y_{min} - Y(x), 0) \tag{9}$$

The mathematical expectation of I(x) can be obtained as follows:

$$E[I(x)] = \left(ymin - \hat{y}(x)\right)\emptyset\left(\frac{y_{min} - \hat{y}(x)}{s}\right) + s\varphi\left(\left(\frac{y_{min} - \hat{y}(x)}{s}\right)\right) \tag{10}$$

where *φ* function composites standard normal cumulative distribution, φ is the probability density of the standard normal distribution function. Additionally, *s* is the standard deviation of the generated agent model. Zhang *et al.,* [23] propose a multipeak parallel adaptive infilling (MPEI) strategy based on expected improvement (EI), which can be divided into two stages: the construction of candidate peak areas and the selection of appropriate candidates at the candidate peak areas. A researcher also suggests implementing mPSO to search for the optimal points that maximize the EI criterion, leading to more efficient and effective optimization.

Infill sampling using the expected improvement criterion with PSO is a powerful approach for global optimization. The EI criterion, in combination with Kriging metamodels, allows for efficient exploration of the search space and identification of promising points for evaluation. PSO further

enhances the optimization process by efficiently searching for the optimal points that maximize the EI criterion.

2.3.4 Greywolf algorithm

Grey Wolf Optimization (GWO) was introduced by Mirjalili *et al.,* [24] and drew inspiration from Particle Swarm Optimization, a well-known metaheuristic method. Previous researchers like Zhang *et al.,* [23] utilized Particle Swarm Optimization to optimize the method of expected improvement. To enhance the performance of the Expected Improvement (EI), this paper adopts the GWO method for optimization. The GWO algorithm replicates the leadership hierarchy and hunting behaviour of grey wolves in nature. Specifically, four types of grey wolves, namely alpha, beta, delta and omega, are employed to simulate the leadership hierarchy. Moreover, the three main hunting steps, which involve searching for prey, encircling prey and attacking prey, are implemented as part of the GWO process.

The GWO algorithm is as follows:

```
Step1: Randomly initialize the Grey wolf population of N particles Xi ( i=1, 2, …, n)
Step2: Calculate the fitness value of each individual
    sort grey wolf population based on fitness values
    alpha_wolf = wolf with the least fitness value
    beta_wolf = wolf with second least fitness value
    gamma_wolf = wolf with third least fitness value
Step 3: For Iter in range(max_iter): # loop max_iter times
    calculate the value of a
      a = 2*(1 - Iter/max_iter)
    For i in range(N): # for each wolf
      a. Compute the value of A1, A2, A3 and C1, C2, C3
         A1 = a*(2*r1 -1), A2 = a*(2*r2 -1), A3 = a*(2*r3 -1)
         C1 = 2*r1, C2 = 2*r2, C3 = 2*r3

      b. Computer X1, X2, X3
         X1 = alpha_wolf.position -
           A1*abs(C1*alpha_wolf_position - ith_wolf.position)
         X2 = beta_wolf.position -
           A2*abs(C2*beta_wolf_position - ith_wolf.position)
         X3 = gamma_wolf.position -
           A3*abs(C3*gamma_wolf_position - ith_wolf.position)

      c. Compute new solution and its fitness
         Xnew = (X1 + X2 + X3) / 3
         fnew = fitness( Xnew)

      d. Update the ith_wolf greedily
        if( fnew < ith_wolf.fitness)
          ith_wolf.position = Xnew
          ith_wolf.fitness = fnew
    End-for
```

```
            compute new alpha, beta and gamma
        sort grey wolf population based on fitness values
        alpha_wolf = wolf with the least fitness value
        beta_wolf = wolf with second least fitness value
        gamma_wolf = wolf with third least fitness value
    End-for
Step 4: Return the best wolf in the population
```

2.4 Proposed Method Flow Chart

The significance of surrogate modelling in optimization problems and its potential to enhance algorithm accuracy is well-documented in a growing body of literature. This section presents the proposed method, outlining how the algorithm predicts input samples and improves the model. Figure 2 shows flow chart for proposed algorithms multi-surrogate with multiple infill sampling. The proposed algorithm is designed to predict two sample points at each iteration, in contrast to the majority of current research work, which predicts only one sample at each iteration to refine the surrogate model.

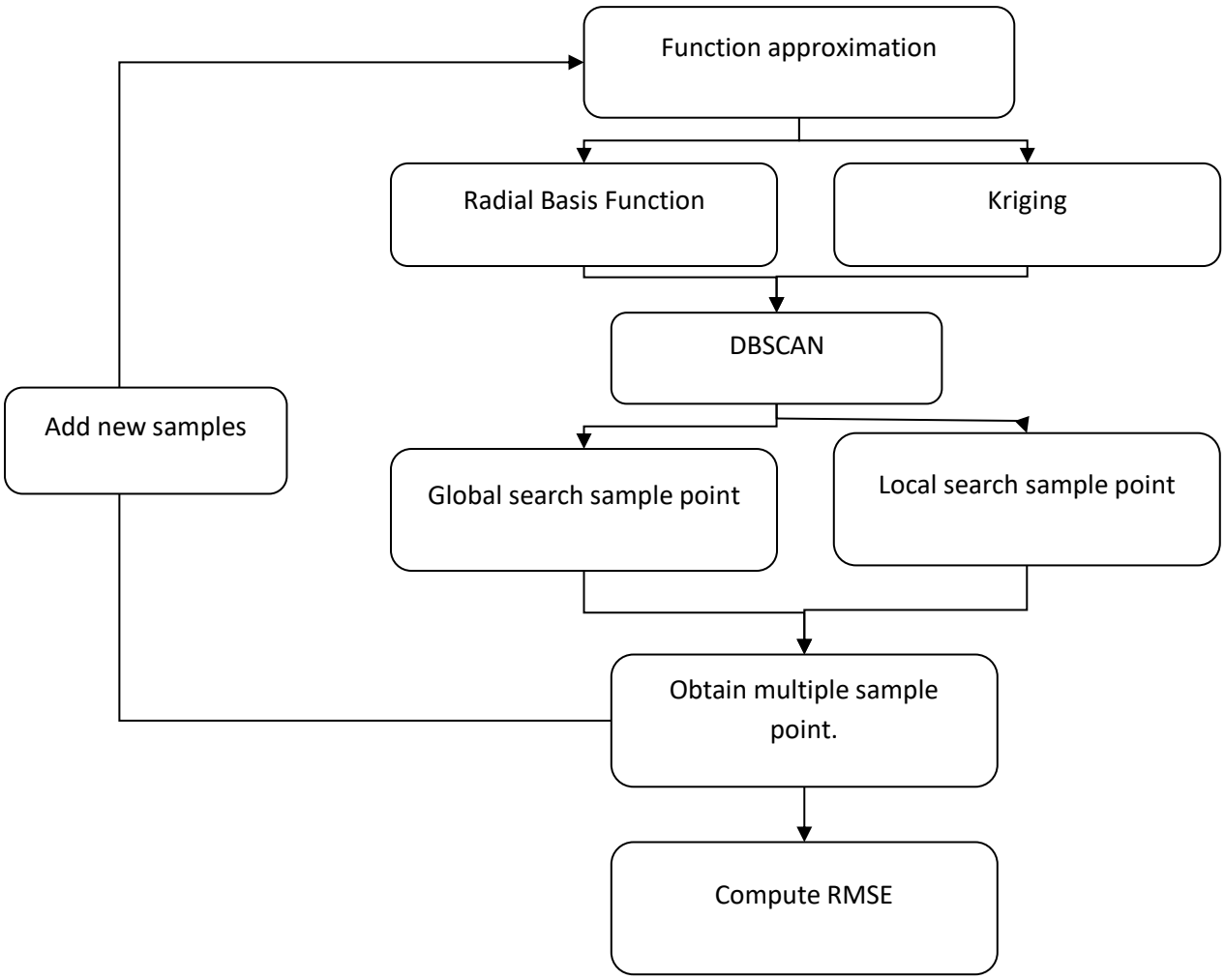

Fig. 2. Flow chart proposed algorithm multi-surrogate multiple infill sampling

In the initial stage, the algorithm utilizes RBF and Kriging as function approximation techniques to predict the optimal sample points of the test function. The subsequent stage employs a clustering algorithm to partition the data into two segments: global search and local search. RBF performs the global search, exploring points across the entire region, while Kriging focuses on the local search, fine-tuning points near the optimal solution. The algorithm carefully selects one sample point for each local and global search, adhering to the designed methodology. Consequently, each iteration yields two sample points, contributing to the refinement of the model through addition to the training data. The global and local search concept aligns with the algorithm's requirement to systematically explore and exploit the experimental region, fostering an efficient and effective optimization process. Table 2 shows the pseudocode of the proposed method multi-surrogate with multiple infill sampling method.

Table 2
Pseudocode of Multi-surrogate Multiple infill sampling

Initial stage:	
Generate sample	
Latin Hypercube Sampling for train data	
Full factorial for test data	
Do:	
DBSN for clustering train data	
Sample near the optimal point – local sample	
Sample far from the optimal point – global sample	
End	
For i: local sample	
Kriging metamodel	Eq. (7) and Eq. (8)
Predict infill sample using EI optimize with GWO	Eq. (9) and Eq. (10) optimize with GWO
End	
For ii: global sample	
RBF metamodel	Eq. (1) - Eq. (5)
Parameter spread and centre using fuzzy cmean	
Predict infill sample using maximin distance	Eq. (6)
End	
Infill 2 samples in each iteration	
Compute RMSE	Eq. (11)
Repeat the algorithm until achieve the stopping criterion	

3. Results

The experimental evidence on the proposed method of multi-surrogate with multiple infill sampling focuses on the low-dimensional test function and the algorithm demonstrate using Modified Easom Function. The actual function of Modified Easom is depicted in Figure 3.

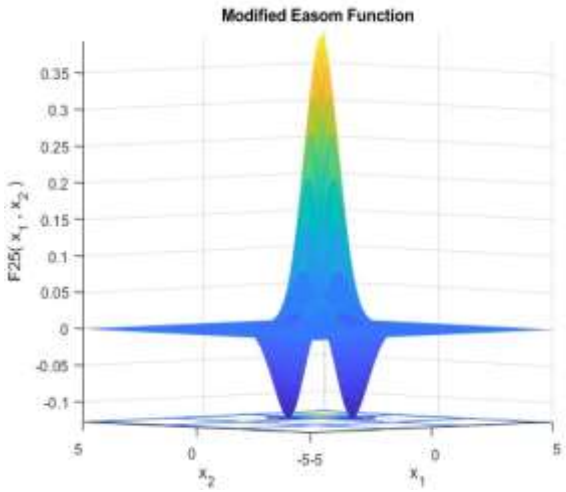

Fig. 3. True function of modified easom

Additionally, Figure 4 displays the initial sample points (in black) and the infill sample points (in red) after 100 sample points were added.

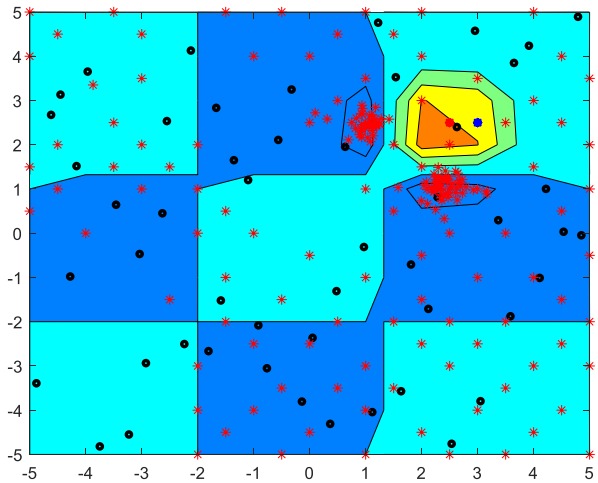

Fig. 4. Contour plot true function after 100 samples added

Figure 5 shows distance between training and test point for algorithm choose the best point as new sample point. The entire computer experiment was completed within approximately 65.16 seconds, involving 50 iterations and the addition of 100 sample points.

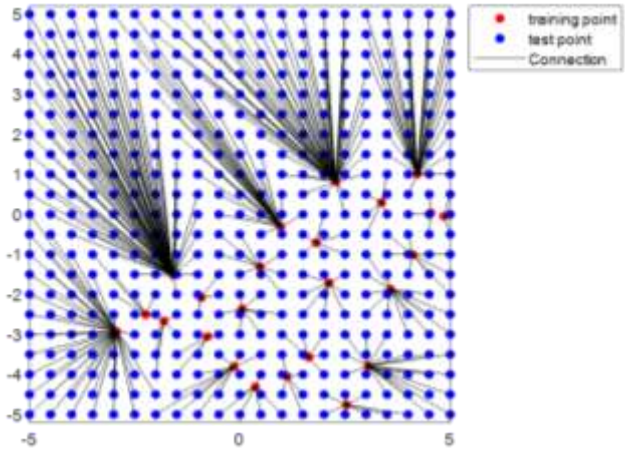

Fig. 5. Train and test sample points

The distribution of training and testing points was investigated using the DBSCAN algorithm, effectively clustering the training samples into global and local sets, as depicted in Figure 6. Notably, local sample points are near the optimal point, while global sample points are positioned far from the optimal points. The global sample relies on the RBF metamodel with maximin distance to predict new sample points. In contrast, the local sample employs the Kriging metamodel, optimized using the Grey Wolf Optimization (GWO) algorithm with the Expected Improvement approach.

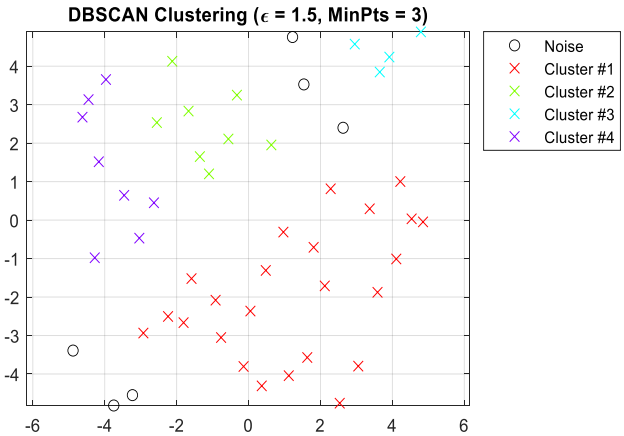

Fig. 6. DBSCAN algorithm clustering the sample point at early stage

The experiment continued for 50 iterations, predicting 100 new infill sample points. The Root Mean Square Error (RMSE) results are illustrated in Figure 7, providing valuable insights into the performance and accuracy. of the multi-surrogate approach with multiple infill sampling.

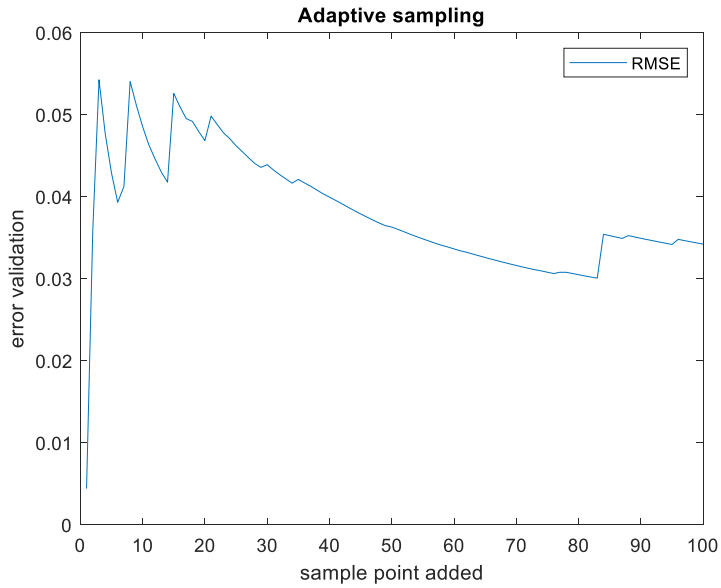

Fig. 7. RMSE validation error after 100 sample points added

Multi-surrogate Multiple infill sampling (MSMIS) was compared with several methods from previous case studies to test and compare algorithm performance. The previous case studies to compare the RMSE performance is SCG (ensemble of surrogate using sign-based cross-validation error with global correction) [25], kriging sequential sampling, RBF sequential sampling and MSMIS. This experiment run to predicts 50 samples point to refine the model and evaluate the accuracy performance. Comparison of result RMSE between previous method and proposed method evaluate using root mean square error (RMSE). Based on numerical comparison of RMSE from previous work extract from other researcher in Table 3, shows that proposed method MSMIS improve the value of RMSE better than one sample for each iteration. This proposed of algorithm is purposely to improve prediction of multiple point at one time. For validation and testing development of algorithm, four (4) benchmarks test function with low dimension was used. Based on the comparison of RMSE, Table 3 shows that the proposed method, MSMIS improves the value of RMSE.

Table 3
Comparison of the performance of each model using RMSE

Test Function	SCG	Kriging	RBF	MSMIS
Goldstein & Price	96495.1	98121.9	100356.9	8988.2
Camelback	1.51710	1.6473	6.3551	1.310
Extended Rosenbrock	2.25e+5	4.46e+e5	2.30e+5	1.99e+5
Branin – Hoo	0.3838	0.3659	9.4963	0.280

3.1 Measuring Accuracy of Metamodel

Two standard performance metrics are used to evaluate the overall performance of the surrogates:

i. Root Mean Squared Error (RMSE), which provides a global error measure across the entire design domain
ii. Maximum Absolute Error (MAE), which represents local deviations.

To compare the performance of various methods across functions, we normalize the RMSE and MAE measures based on the actual function values [26,27]. A set of data of the size known as *nt* and a set of predictions at those locations and calculate the RMSE:

$$RMSE = \sqrt{\frac{1}{nt}\sum_{k=1}^{nt}\left(f(x^k) - \tilde{f}(x^k)\right)^2} \tag{11}$$

where $f(x^k)$ represents the exact function value for the test point x^k, $\tilde{f}(x^k)$is the corresponding estimated function value and nt is the number of test points chosen for evaluating the error measure. Generally, the RMSE metric should be as small as possible.

4. Conclusions

This work introduces an innovative approach, the multi-surrogates and multi-points infill global optimization method, designed to tackle computational time for black-box optimization problems effectively. A multi-point infill strategy is proposed to address the issue of sparsely sampled regions, which iteratively adds multiple approximate optimal points and promising points. The pre-processing stage involves implementing the DBSCAN algorithm to partition the training data before function approximation and predict new sample points. The algorithm is thoroughly tested using four (4) popular mathematical test functions, drawing on insights from previous related studies. Remarkably, the proposed method significantly enhances the accuracy of the surrogate model, as evidenced by experimental results. The algorithm should be tested against real-world optimization problems for future endeavours, showcasing its potential to contribute to practical applications.

Acknowledgment
This research not funded by any grant but Ungku Omar Polytechnic, Ministry of Higher Education, funded this research paper for publication.

References
[1] Ouyang, Qi, Wenxi Lu, Tiansheng Miao, Wenbing Deng, Changlong Jiang and Jiannan Luo. "Application of ensemble surrogates and adaptive sequential sampling to optimal groundwater remediation design at DNAPLs-contaminated sites." *Journal of contaminant hydrology* 207 (2017): 31-38. https://doi.org/10.1016/j.jconhyd.2017.10.007
[2] Zhou, Qi, Yan Wang, Ping Jiang, Xinyu Shao, Seung-Kyum Choi, Jiexiang Hu, Longchao Cao and Xiangzheng Meng. "An active learning radial basis function modeling method based on self-organization maps for simulation-based design problems." *Knowledge-Based Systems* 131 (2017): 10-27. https://doi.org/10.1016/j.knosys.2017.05.025
[3] Durantin, Cédric, Justin Rouxel, Jean-Antoine Désidéri and Alain Glière. "Multifidelity surrogate modeling based on radial basis functions." *Structural and Multidisciplinary Optimization* 56 (2017): 1061-1075. https://doi.org/10.1007/s00158-017-1703-7
[4] Aburashed, Laila, Marah AL Amoush and Wardeh Alrefai. "SQL Injection Attack Detection using Machine Learning Algorithms." *Semarak International Journal of Machine Learning* 2, no. 1 (2024): 1-12. https://doi.org/10.37934/sijml.2.1.112
[5] Wang, Yuan, Zhong-Hua Han, Yu Zhang and Wen-Ping Song. "Efficient global optimization using multiple infill sampling criteria and surrogate models." In *2018 AIAA Aerospace Sciences Meeting*, p. 0555. 2018. https://doi.org/10.2514/6.2018-0555
[6] Chao, S. O. N. G., Y. A. N. G. Xudong and S. O. N. G. Wenping. "Multi-infill strategy for kriging models used in variable fidelity optimization." *Chinese Journal of Aeronautics* 31, no. 3 (2018): 448-456. https://doi.org/10.1016/j.cja.2018.01.011
[7] Song, Xiaodong, Mingyang Li, Zhitao Li and Fang Liu. "Global Optimization Algorithm Based on Kriging Using Multi-Point Infill Sampling Criterion and Its Application in Transportation System." *Sustainability* 13, no. 19 (2021): 10645. https://doi.org/10.3390/su131910645

[8] Hwang, Yongmoon, Sang-Lyul Cha, Sehoon Kim, Seung-Seop Jin and Hyung-Jo Jung. "The multiple-update-infill sampling method using minimum energy design for sequential surrogate modeling." *Applied Sciences* 8, no. 4 (2018): 481. https://doi.org/10.3390/app8040481

[9] Qin, Shufen, Chan Li, Chaoli Sun, Guochen Zhang and Xiaobo Li. "Multiple infill criterion-assisted hybrid evolutionary optimization for medium-dimensional computationally expensive problems." *Complex & Intelligent Systems* (2021): 1-13. https://doi.org/10.1007/s40747-021-00541-4

[10] Yu, Mingyuan, Jing Liang, Kai Zhao and Zhou Wu. "An aRBF surrogate-assisted neighborhood field optimizer for expensive problems." *Swarm and Evolutionary Computation* 68 (2022): 100972. https://doi.org/10.1016/j.swevo.2021.100972

[11] Habib, Ahsanul, Hemant Kumar Singh and Tapabrata Ray. "A multi-objective batch infill strategy for efficient global optimization." In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 4336-4343. IEEE, 2016. https://doi.org/10.1109/CEC.2016.7744341

[12] Fico, Francesco, Francesco Urbino, Robert Carrese, Pier Marzocca and Xiaodong Li. "Surrogate-assisted multi-swarm particle swarm optimization of morphing airfoils." In *Artificial Life and Computational Intelligence: Third Australasian Conference, ACALCI 2017, Geelong, VIC, Australia, January 31–February 2, 2017, Proceedings 3*, pp. 124-133. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-51691-2_11

[13] Chevalier, Clément, Victor Picheny and David Ginsbourger. "Kriginv: An efficient and user-friendly implementation of batch-sequential inversion strategies based on kriging." *Computational statistics & data analysis* 71 (2014): 1021-1034. https://doi.org/10.1016/j.csda.2013.03.008

[14] Xing, Jian, Yangjun Luo and Zhonghao Gao. "A global optimization strategy based on the Kriging surrogate model and parallel computing." *Structural and Multidisciplinary Optimization* 62 (2020): 405-417. https://doi.org/10.1007/s00158-020-02495-6

[15] Chen, Cong, Jiaxin Liu and Pingfei Xu. "Comparison of parallel infill sampling criteria based on Kriging surrogate model." *Scientific Reports* 12, no. 1 (2022): 678. https://doi.org/10.1038/s41598-021-04553-5

[16] Yang, Kaifeng, Michael Affenzeller and Guozhi Dong. "A parallel technique for multi-objective Bayesian global optimization: Using a batch selection of probability of improvement." *Swarm and evolutionary computation* 75 (2022): 101183. https://doi.org/10.1016/j.swevo.2022.101183

[17] Haftka, Raphael T., Diane Villanueva and Anirban Chaudhuri. "Parallel surrogate-assisted global optimization with expensive functions–a survey." *Structural and Multidisciplinary Optimization* 54 (2016): 3-13. https://doi.org/10.1007/s00158-016-1432-3

[18] Lin, Dennis KJ, Timothy W. Simpson and Wei Chen. "Sampling strategies for computer experiments: design and analysis." *International Journal of Reliability and applications* 2, no. 3 (2001): 209-240.

[19] Daszykowski, M. and B. Walczak. "Density-based clustering methods." *Comprehensive Chemometrics* (2009): 635-654. https://doi.org/10.1016/B978-044452701-1.00067-3

[20] Kremers, Bart JJ, Jonathan Citrin, Aaron Ho and Karel L. van de Plassche. "Two-step clustering for data reduction combining DBSCAN and k-means clustering." *Contributions to Plasma Physics* 63, no. 5-6 (2023): e202200177. https://doi.org/10.1002/ctpp.202200177

[21] Dash, Ch Sanjeev Kumar, Ajit Kumar Behera, Satchidananda Dehuri and Sung-Bae Cho. "Radial basis function neural networks: a topical state-of-the-art survey." *Open Computer Science* 6, no. 1 (2016): 33-63. https://doi.org/10.1515/comp-2016-0005

[22] Xi, Rui, Hongguang Jia and Qianjin Xiao. "Study of experimental design and response surface method for surrogate model of computational simulation." In *2011 International Conference on Electrical and Control Engineering*, pp. 4995-4998. IEEE, 2011. https://doi.org/10.1109/ICECENG.2011.6057240

[23] Zhang, Yang, Shuo Wang, Chang'an Zhou, Liye Lv and Xueguan Song. "A fast active learning method in design of experiments: multipeak parallel adaptive infilling strategy based on expected improvement." *Structural and Multidisciplinary Optimization* 64, no. 3 (2021): 1259-1284. https://doi.org/10.1007/s00158-021-02915-1

[24] Mirjalili, Seyedali, Seyed Mohammad Mirjalili and Andrew Lewis. "Grey wolf optimizer." *Advances in engineering software* 69 (2014): 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007

[25] Qiu, Haobo, Liming Chen, Chen Jiang, Xiwen Cai and Liang Gao. "Ensemble of surrogate models using sign based cross validation error." In *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 526-531. IEEE, 2017. https://doi.org/10.1109/CSCWD.2017.8066749

[26] Liu, Haitao, Shengli Xu and Xiaofang Wang. "Sequential sampling designs based on space reduction." *Engineering Optimization* 47, no. 7 (2015): 867-884. https://doi.org/10.1080/0305215X.2014.928816

[27] Ye, Pengcheng and Guang Pan. "Selecting the best quantity and variety of surrogates for an ensemble model." *Mathematics* 8, no. 10 (2020): 1721. https://doi.org/10.3390/math8101721