Metaheuristic Optimization based Feature Selection for Software Defect Prediction

Romi Satria Wahono
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka
Graduate School of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
Email: romi@brainmatics.com

Nanna Suryana and Sabrina Ahmad
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka
Email: {nsuryana, sabrinaahmad}@utem.edu.my

***Abstract*—Software defect prediction has been an important research topic in the software engineering field, especially to solve the inefficiency and ineffectiveness of existing industrial approach of software testing and reviews. The software defect prediction performance decreases significantly because the data set contains noisy attributes and class imbalance. Feature selection is generally used in machine learning when the learning task involves high-dimensional and noisy attribute datasets. Most of the feature selection algorithms, use local search throughout the entire process, consequently near-optimal to optimal solutions are quiet difficult to be achieved. Metaheuristic optimization can find a solution in the full search space and use a global search ability, significantly increasing the ability of finding high-quality solutions within a reasonable period of time. In this research, we propose the combination of metaheuristic optimization methods and bagging technique for improving the performance of the software defect prediction. Metaherustic optimization methods (genetic algorithm and particle swarm optimization) are applied to deal with the feature selection, and bagging technique is employed to deal with the class imbalance problem. Results have indicated that the proposed methods makes an impressive improvement in prediction performance for most classifiers. Based on the comparison result, we conclude that there is no significant difference between particle swarm optimization and genetic algorithm when used as feature selection for most classifiers in software defect prediction.**

***Index Terms*—software defect prediction, feature selection, genetic algorithm, particle swarm optimization, bagging technique**

I. Introduction

Software defects or software faults are expensive in quality and cost. It is a deficiency in a software product that causes it to perform unexpectedly [1]. Moreover, the cost of capturing and correcting defects is one of the most expensive software development activities [2]. The accurate prediction of defect-prone software modules can help direct test effort, reduce costs, improve the quality of software [3], reach a highly dependable system, improving the test process by focusing on fault-prone modules, and identifying refactoring candidates that are predicted as fault- prone [4].

Recent studies showed that the probability of detection of fault prediction models might be higher than the probability of detection of software reviews. Menzies et al. found defect predictors with a probability of detection of 71 percent [5]. This is markedly higher than other currently used industrial methods such as manual code reviews. A panel at IEEE Metrics 2002 [6] concluded that manual software reviews can find 60 percent of defects. Therefore, software fault prediction approaches are much more efficient and effective to detect software faults compared to software reviews.

Software defect prediction has been an important research topic in the software engineering field [3]. Current software defect prediction research focuses on three topics [7]. First topic concern on estimating the number of defects remaining in software systems, the second one, concern on discovering defect associations, and the third topic, focus on classifying the defect-proneness of software components, typically into two classes, defect-prone and not defect-prone.

The first type of work employs statistical approaches [8] [9], capture-recapture models [10] [11] [12], and detection profile methods [13] to estimate the number of defects remaining in software systems with inspection data and process quality data. The prediction result can be used as an important measure for the software developer and can be used to control the software process and gauge the likely delivered quality of a software system [14].

The second type of work borrows association rule mining algorithms from the data mining community to reveal software defect associations [15] [16] [17] which can be used for three purposes. First, finding as many related defects as possible to the detected defects and consequently, make more effective corrections to the software. Second, helping to evaluate reviewers' results during an inspection. Thus, a recommendation might be that their work should be reinspected for completeness. Third, assisting managers in improving the software process through analysis of the reasons why some defects frequently occur together. If the analysis leads to the

doi:10.4304/jsw.9.5.1324-1333

identification of a process problem, managers can devise corrective action.

The third type of work classifies software components or modules as defect-prone and non-defect-prone by means of metric based classification [18] [19] [20] [21] [7]. Classification algorithm is a popular machine learning approach for software defect prediction [21]. It categorizes the software code attributes into defective or not defective, which is completed by means of a classification model derived from software metrics data from previous development projects [22]. Being able to predict which components are more likely to be defect-prone supports better targeted testing resources and therefore, improved efficiency. This research is focused and concerned with the third approach.

Various types of classification algorithms have been applied for software defect prediction, including logistic regression [23], decision trees [24], neural networks [25], naïve-bayes [20]. Unfortunately, software defect prediction remains a largely unsolved problem. The comparisons and benchmarking result of the defect prediction using machine learning classifiers indicate that, no significant performance differences could be detected [21] and no particular classifiers that performs the best for all the data sets [7]. There is a need of accurate defect prediction model for large-scale software system.

Two common aspects of data quality that can affect classification performance are class imbalance and noisy attributes [26] of data sets. Software defect datasets have an imbalanced nature with very few defective modules compared to defect-free ones [27]. Imbalance can lead to a model that is not practical in software defect prediction, because most instances will be predicted as non-defect prone [28]. Learning from imbalanced datasets is difficult. The insufficient information that is associated with the minority class impedes making a clear understanding of the inherent structure of the dataset [29]. The software defect prediction performance also decreases significantly because the dataset contains noisy attributes [30] [31]. However, the noisy data points in the datasets that cannot be confidently assumed to be erroneous using such simple method [32].

Feature selection is generally used in machine learning when the learning task involves high-dimensional and noisy attribute datasets. Most of the feature selection algorithms, use local search throughout the entire process, consequently near-optimal to optimal solutions are quiet difficult to be achieved. Metaheuristic optimization can find a solution in the full search space and use a global search ability, significantly increasing the ability of finding high-quality solutions within a reasonable period of time [33]. Mostly used metaheuristic optimization for feature selection includes genetic algorithm (GA), particle swarm optimization (PSO) and ant colony optimization (ACO).

In the current work, we propose the combination of metaheuristic optimization methods (GA and PSO) and bagging technique for improving the accuracy of software defect prediction. Metaheuristic optimization methods are applied to deal with the feature selection, and bagging technique is employed to deal with the class imbalance problem. Bagging technique is chosen due to the effectiveness in handling class imbalance [26]. The proposed method is evaluated using the state-of-the-art and public datasets from NASA metric data repository.

II. Related Research

Feature selection is an important data preprocessing activity and has been extensively studied in the data mining and machine learning community. The main goal of feature selection is to select a subset of features that minimizes the prediction errors of classifiers. Feature selection techniques are divided into two categories: wrapper-based approach and filter-based approach. The wrapper-based approach involves training a learner during the feature selection process, while the filter-based approach uses the intrinsic characteristics of the data, based on a given metric, for feature selection and does not depend on training a learner. The primary advantage of the filter-based approach over the wrapper-based approach is that it is computationally faster. However, if computational complexity was not a factor, then a wrapper-based approach was the best overall feature selection scheme in terms of accuracy. Because the objective of this research is to improve the quality and accuracy of software defect prediction model, it was decided to use the wrapper-based approach. Nevertheless, wrapper methods have the associated problem of having to train a classifier for each tested feature subset. This means testing all the possible combinations of features will be virtually impossible.

Most of the feature selection strategies attempt to find solutions that range between sub-optimal and near optimal regions. They use local search throughout the entire process, instead of global search. On the other hand, these search algorithms utilize a partial search over the feature space, and suffer from computational complexity. Consequently, near-optimal to optimal solutions are quiet difficult to achieve using these algorithms. As a result, many research studies now focus on metaheuristic optimization techniques [34]. The significance of metaheuristic optimization techniques is that they can find a solution in the full search space on the basis of activities of multi-agent systems that use a global search ability utilizing local search appropriately, thus significantly increasing the ability of finding very high-quality solutions within a reasonable period of time [33]. Metaheuristic optimization techniques have been developed in several domains and include algorithms like simulated annealing, tabu-search, as well as bio-inspired methods like genetic algorithms, evolution strategies, ant colony optimization and particle swarm optimization. These methods are able to find fairly good solutions without searching the entire workspace.

Although feature selection has been widely applied in numerous application domains for many years, its application in the software quality prediction domain is limited. Song et al. [7] applied two wrapper approaches, forward selection and backward elimination, as a feature selection for their proposed model. Song et al. concluded

that a feature selection techniques, especially forward selection and backward elimination can play different roles with different learning algorithms for different data sets and that no learning scheme dominates, i.e., always outperforms the others for all data sets. This means we should choose different learning schemes for different data sets, and consequently, the evaluation and decision process is important. Wang et al. [35] applied ensemble feature selection techniques to 16 software defect data sets, and they concluded that ensembles of very few rankers are very effective and even better than ensembles of many or all rankers.

The class imbalance problem is observed in various domain, including software defect prediction. Several methods have been proposed in literature to deal with class imbalance: data sampling, boosting and bagging. Data sampling is the primary approach for handling class imbalance, and it involves balancing the relative class distributions of the given data set. There are two types of data sampling approaches: under sampling and oversampling [36]. Boosting is another technique, which is very effective when learning from imbalanced data. Compared to data sampling, boosting has received relatively little attention in data-mining research with respect to class imbalance. However, Seiffert et al. [36] show that boosting performs very well. Bagging, may outperform boosting when data contain noise [37], because boosting may attempt to build models to correctly classify noisy examples. In this study we apply bagging technique, because Khoshgoftaar et al. [26] showed that the bagging techniques generally outperform boosting, and hence in noisy data environments. Therefore bagging is the preferred method for handling class imbalance.

While considerable work has been done for feature selection and class imbalance problem separately, limited research can be found on investigating them both together, particularly in the software engineering field [26]. In this study, we combine metaheuristic optimization methods (GA and PSO) for selecting features and bagging technique for solving the class imbalance problem, in the context of software defect prediction.

III. Proposed Software Defect Prediction Framework

A. Integration of GA Based Feature Selection Method and Bagging Technique

Figure1 shows an activity diagram of the integration of Bagging technique and GA based feature selection. The aim of GA is to find optimum solution within the potential solution set. Each solution set is called as population. Populations are composed of vectors, namely, chromosome or individual. Each item in the vector is called as gene. In the proposed method, chromosomes represent features, which are encoded as binary strings of 1 and 0. In this scheme, 1 represents se- lection of a feature and 0 means a non-selection.

As shown in Figure 1, input data set includes training data set and testing data set. Relational feature subsets are chosen and unrelated features subsets are discarded by feature subset selection. After training data set and testing data set discarded unrelated feature subsets, they become training data set of selected feature subset and testing data set of selected feature subset. Classifiers are trained by training set with selected feature subset.

Bagging (Bootstrap Aggregating) was proposed by Leo Breiman in 1994 [38] to improve the classification by combining classifications of randomly generated training sets. The bagging classifier separates a training set into several new training sets by random sampling, and builds models based on the new training sets. The final classification result is obtained by the voting of each model. It also reduces variance and helps to avoid overfitting. Description of the bagging technique is as follows. Given a standard training set D of size n, bagging generates m new training sets D_i, each of size $n'<n$, by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If $n'=n$, then for large n the set D_i is expected to have the fraction $(1 - 1/e)$ of the unique examples of D, the rest being duplicates. This kind of sample is known as a bootstrap sample. The m models are fitted using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification). Bagging leads to improvements for unstable procedures [38], which include neural network, classification and regression trees, and subset selection in linear regression. On the other hand, it can mildly degrade the performance of stable methods such as K-nearest neighbors.

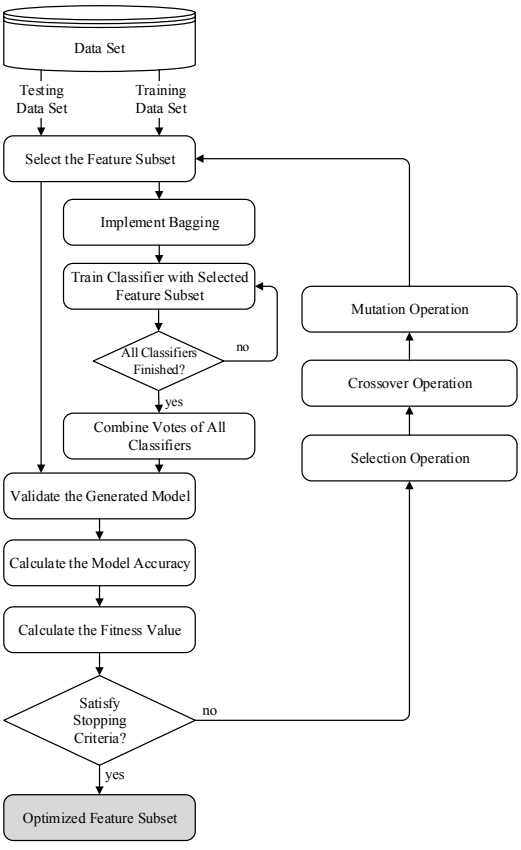

Figure 1. Activity Diagram of the Integration of Bagging Technique and Genetic Algorithm based Feature Selection

Classification accuracy of classifier is calculated by testing set with selected feature subset. Classification accuracy of classifier, the number of selected features and the feature cost are used to construct a fitness function. Every chromosome is evaluated by the equation (1).

$$fitness = W_A \times A + W_F \times \left(P + \left(\sum_{i=1}^{n_f} C_i \times F_i\right)\right)^{-1} \quad (1)$$

where A is classification accuracy, W_A is weight of classification accuracy, F_i is feature value, W_F is feature weight, C_i is feature cost, P is setting constant of avoiding that denominator reaches zero.

When ending condition is satisfied, the operation ends, otherwise, continue with the next generation operation. The proposed method searches for better solutions by genetic operations, including crossover, mutation and selection.

B. Integration of PSO based Feature Selection and Bagging Technique

Particle swarm optimization (PSO) is an emerging population-based meta-heuristic that simulates social behavior such as birds flocking to a promising position to achieve precise objectives in a multi-dimensional space. PSO performs searches using a population (*swarm*) of individuals (*particles*) that are updated from iteration to iteration. The size of population is denoted as p_{size}. To discover the optimal solution, each particle changes its searching direction according to two factors, its own best previous experience (*pbest*) and the best experience of all other members (*gbest*). Shi and Eberhart [39] called *pbest* the cognition part, and *gbest* the social part.

Each particle represents a candidate position (solution). A particle is considered as a point in a *D*-dimension space, and its status is characterized according to its position and velocity. The *D*-dimensional position for the particle i at iteration t can be represented as $x_i^t = \{x_{i1}^t, x_{i2}^t, \dots, x_{iD}^t\}$. Likewise, the velocity (distance charge) for particle i at iteration t, which is also a *D*-dimension vector, can be described as $v_i^t = \{v_{i1}^t, v_{i2}^t, \dots, v_{iD}^t\}$.

In the later version of PSO, a new parameter, called inertia weight introduced by [39] due to control over the previous velocity of the particles. Let $p_i^t = \{p_{i1}^t, p_{i2}^t, \dots, p_{iD}^t\}$ represent the best solution that particle i has obtained until iteration t, and $p_g^t = \{p_{g1}^t, p_{g2}^t, \dots, p_{gD}^t\}$ denote the best solution obtained from p_i^t in the population at iteration t. To search for the optimal solution, each particle changes its velocity according to the cognitive and social part using equation (2).

$$V_{id}^t = w * V_{id}^{t-1} + c_1 r_1 (P_{id}^t - x_{id}^t) + c_2 r_2 (P_{gd}^t - x_{id}^t) \quad (2)$$

Note that, c_1 indicates the cognitive learning factor, c_2 indicates the social learning factor, inertia weight (w) is used to slowly reduce the velocity of the particles to keep the swarm under control, and r_1 and r_2 are random numbers uniformly distributed in *U(0,1)*.

Each particle then moves to a new potential solution based on the equation (3).

$$X_{id}^{t+1} = X_{id}^t + V_{id}^t \quad (3)$$

Figure 2 shows an activity diagram of the integration of bagging technique and particle swarm optimization based feature selection. A group of particles are random generated, dimensional discrete binary variable. The particle length is the total characteristics number, and if and each particle is one the first *i*-bit is 1, then the first feature *i* was selected, otherwise it will be shielded. Each particle represents a feature subset, which is a candidate solution. Implement bagging technique and train the classifier on the larger training set based on the selected feature subset and the type of kernel. If all classifiers are finished, combine votes of all classifiers. Finally, measure validation accuracy on testing data set via the generated model.

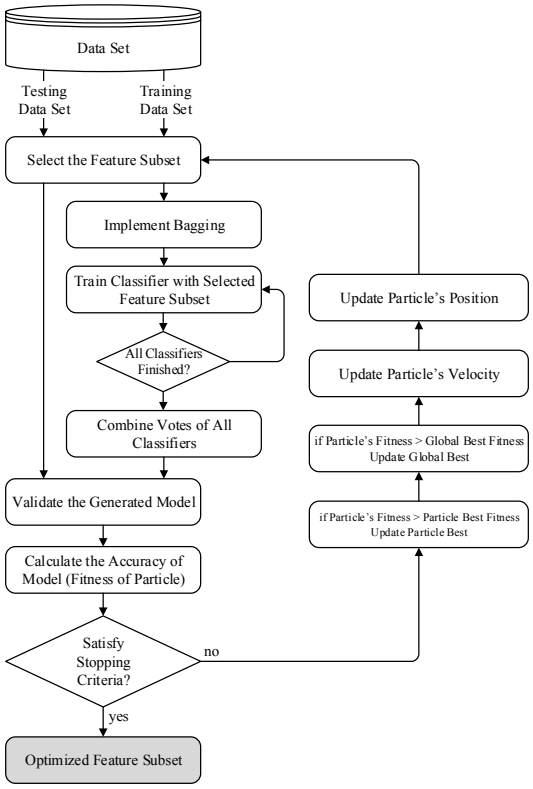

Figure 2. Activity Diagram of the Integration of Bagging Technique and Particle Swarm Optimization based Feature Selection

We use the accuracy of classifier to evaluate fitness size, the higher accuracy rate, the greater the fitness. the goal of select the feature characteristic subset that is to the achieve the use a small number of same or better classification results, so the fitness function evaluation should also take into consideration the number of characteristics, given same accuracy to two characteristics of a subset, the one which have fewer characteristic number will be higher fitness.

Update the global and personal best according to the fitness evaluation results. Record the average validation

accuracy for the global and personal best. If the fitness is better than the particle's best fitness, then the position vector is saved for the particle. If the particle's fitness is better than the global best fitness, then the position vector is saved for the global best. Finally the particle's velocity and position are updated until the termination condition is satisfied. To avoid overtraining, we observe the validation accuracy curve, and stop training when the iteration has the best validation accuracy during the training process. With the stopping training iteration determined in the previous step, recall the recorded feature subset and the type of kernel in the stopping iteration.

IV. Experimental Setup

A. Framework

The framework of our experiment is shown in Figure 3. The framework is comprised 1) a data sets 2) a feature selection, 3) a meta-learning, 4) a learning algorithm, 5) a model validation, 6) a model evaluation and 7) a model comparison. The used platform is Intel Core i7 2.2 GHz CPU, 16 GB RAM, and Microsoft Windows 7 Professional 64-bit with SP1 operating system. The development environment is Netbeans 7 with Java programming language. The application software is RapidMiner 5.2.

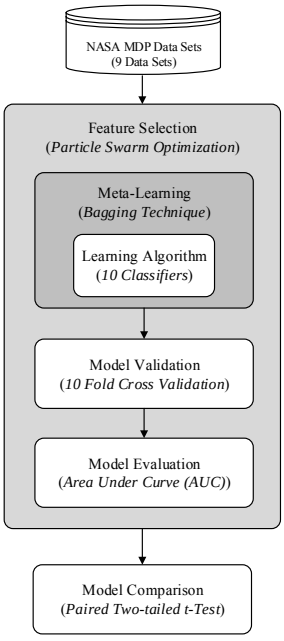

Figure 3. The Framework of Experiment

B. Data Sets

One of the most important problems for software fault prediction studies is the usage of nonpublic (private) data sets. Several companies developed fault prediction models using proprietary data and presented these models in conferences. However, it is not possible to compare results of such studies with results of our own models because their datasets cannot be reached. The use of public datasets makes our research repeatable, refutable, and verifiable [40]. Recently, state-of-the-art public data sets used for software defect prediction research is available in NASA Metrics Data (MDP) repository [32].

The data used in this research are collected from the NASA MDP repository. NASA MDP repository is a database that stores problem, product, and metrics data [32]. The primary goal of this data repository is to provide project data to the software community. In doing so, the Metrics Data Program collects artifacts from a large NASA dataset, generates metrics on the artifacts, and then generates reports that are made available to the public at no cost. The data that are made available to general users have been sanitized and authorized for publication through the Metrics Data Program Web site by officials representing the projects from which the data originated.

TABLE I.
NASA MDP Data Sets and The Code Attributes

Code Attributes		NASA MDP dataset								
		CM1	KC1	KC3	MC2	MW1	PC1	PC2	PC3	PC4
LOC counts	LOC_total	√	√	√	√	√	√	√	√	√
	LOC_blank	√	√	√	√	√	√	√	√	√
	LOC_code_and_comment	√	√	√	√	√	√	√	√	√
	LOC_comments	√	√	√	√	√	√	√	√	√
	LOC_executable	√	√	√	√	√	√	√	√	√
	number_of_lines	√		√	√	√	√	√	√	√
Halstead	content	√	√	√	√	√	√	√	√	√
	difficulty	√	√	√	√	√	√	√	√	√
	effort	√	√	√	√	√	√	√	√	√
	error_est	√	√	√	√	√	√	√	√	√
	length	√	√	√	√	√	√	√	√	√
	level	√	√	√	√	√	√	√	√	√
	prog_time	√	√	√	√	√	√	√	√	√
	volume	√	√	√	√	√	√	√	√	√
	num_operands	√	√	√	√	√	√	√	√	√
	num_operators	√	√	√	√	√	√	√	√	√
	num_unique_operands	√	√	√	√	√	√	√	√	√
	num_unique_operators	√	√	√	√	√	√	√	√	√
McCabe	cyclomatic_complexity	√	√	√	√	√	√	√	√	√
	cyclomatic_density	√		√	√	√	√	√	√	√
	design_complexity	√	√	√	√	√	√	√	√	√
	essential_complexity	√	√	√	√	√	√	√	√	√
Misc.	branch_count	√	√	√	√	√	√	√	√	√
	call_pairs	√		√	√	√	√	√	√	√
	condition_count	√		√	√	√	√	√	√	√
	decision_count	√		√	√	√	√	√	√	√
	decision_density	√		√	√	√	√	√	√	√
	edge_count	√		√	√	√	√	√	√	√
	essential_density	√		√	√	√	√	√	√	√
	parameter_count	√		√	√	√	√	√	√	√
	maintenance_severity	√		√	√	√	√	√	√	√
	modified_condition_count	√		√	√	√	√	√	√	√
	multiple_condition_count	√		√	√	√	√	√	√	√
	global_data_complexity			√	√					
	global_data_density			√	√					
	normalized_cyclomatic_complexity	√		√	√	√	√	√	√	√
	percent_comments	√		√	√	√	√	√	√	√
	node_count	√		√	√	√	√	√	√	√
Programming Language		*C*	*C++*	*Java*	*C*	*C*	*C*	*C*	*C*	*C*
Number of Code Attributes		*37*	*21*	*39*	*39*	*37*	*37*	*77*	*37*	*37*
Number of Modules		*505*	*1571*	*458*	*127*	*403*	*1059*	*4505*	*1511*	*1347*
Number of fp Modules		*48*	*319*	*43*	*44*	*31*	*76*	*23*	*160*	*178*
Percentage of fp Modules		*9.5*	*20.31*	*9.39*	*34.65*	*7.69*	*7.18*	*0.51*	*10.59*	*13.21*

Each NASA data set is comprised of several software modules, together with their number of faults and characteristic code attributes. After preprocessing, modules that contain one or more errors were labeled as fault-prone, whereas error-free modules were categorized as not-fault-prone. Besides line of codes (LOC) counts, the NASA MDP data sets include several Halstead attributes as well as McCabe complexity measures. The former estimates reading complexity by counting operators and operands in a module, whereas the latter is derived from a module's flow graph.

Some researchers endorse the static code attributes defined by McCabe and Halstead as defect predictors in the software defect prediction. McCabe and Halstead are module-based metrics, where a module is the smallest unit of functionality. Static code attributes are used as defect predictors, since they are useful, generalizable, easy to use, and widely used [27].

In this research, we use nine software defect prediction data sets from NASA MDP. Individual attributes per data set, together with some general statistics and descriptions, are given in Table I. These data sets have various scales of LOC, various software modules coded by several different programming languages including C, C++ and Java, and various types of code metrics including code size, Halstead's complexity and McCabe's cyclomatic complexity.

C. Model Validation

We use the state-of-the-art stratified 10-fold cross-validation for learning and testing data. This means that we divided the training data into 10 equal parts and then performed the learning process 10 times. Each time, we chose another part of dataset for testing and used the remaining nine parts for learning. After, we calculated the average values and the deviation values from the ten different testing results. We employ the stratified 10-fold cross validation, because this method has become the standard method in practical terms. Some tests have also shown that the use of stratification improves results slightly [41].

D. Model Evaluation

We applied *area under curve* (AUC) as an accuracy indicator in our experiments to evaluate the performance of classifiers. AUC is area under ROC curve. Lessmann et al. [21] advocated the use of the AUC to improve cross-study comparability. The AUC has the potential to significantly improve convergence across empirical experiments in software defect prediction, because it separates predictive performance from operating conditions, and represents a general measure of predictiveness. Furthermore, the AUC has a clear statistical interpretation. It measures the probability that a classifier ranks a randomly chosen fault-prone module higher than a randomly chosen non-fault-prone module. Consequently, any classifier achieving AUC well above 0.5 is demonstrably effective for identifying fault-prone modules and gives valuable advice as to which modules should receive particular attention in software testing.

A rough guide for classifying the accuracy of a diagnostic test using AUC is the traditional system, presented below [42]:

- 0.90 - 1.00 = excellent classification
- 0.80 - 0.90 = good classification
- 0.70 - 0.80 = fair classification
- 0.60 - 0.70 = poor classification
- 0.50 - 0.60 = failure

E. Model Comparison using Paired Two-tailed t-Test

A paired t-test compares two samples in cases where each value in one sample has a natural partner in the other. A paired t-test looks at the difference between paired values in two samples, takes into account the variation of values within each sample, and produces a single number known as a t-value. In this research, we have applied pair wise t-tests over mean values of each datasets in order to determine statistical significance of results with $\propto = 0.05$.

V. Results and Analysis

A. Preliminary Results

First of all, we conducted experiments on 9 NASA MDP data sets by using 10 classification algorithms. More specifically, it applies five types of classification models that include traditional statistical classifiers (Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Naïve Bayes (NB)), Nearest Neighbors (k-nearest neighbor (k-NN) and K*), Neural Network (Back Propagation (BP)), Support Vector Machine (SVM), and Decision Tree (C4.5, Classification and Regression Tree (CART), and Random Forest (RF)).

The experimental results are reported in Table II. This result confirmed Hall et al. result [3] that NB and LR, in particular, seem to be the techniques used in models that are performing relatively well in software defect prediction. Models based on Decision Tree seem to underperform due to the class imbalance. SVM techniques also perform less well, though SVM has excellent generalization ability in the situation of small sample data like NASA MDP data set.

TABLE II.
AUC of 10 Classifiers on 9 Data Sets

Classifiers		CM1	KC1	KC3	MC2	MW1	PC1	PC2	PC3	PC4
Statistical Classifier	LR	0.763	0.801	0.713	0.766	0.726	0.852	0.849	0.81	0.894
	LDA	0.471	0.536	0.447	0.503	0.58	0.454	0.577	0.524	0.61
	NB	0.734	0.786	0.67	0.739	0.732	0.781	0.811	0.756	0.838
Nearest Neighbor	k-NN	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	K*	0.6	0.678	0.562	0.585	0.63	0.652	0.754	0.697	0.76
Neural Network	BP	0.713	0.791	0.647	0.71	0.625	0.784	0.918	0.79	0.883
Support Vector Machine	SVM	0.753	0.752	0.642	0.761	0.714	0.79	0.534	0.75	0.899
Decision Tree	C4.5	0.565	0.515	0.497	0.455	0.543	0.601	0.493	0.715	0.723
	CART	0.604	0.648	0.637	0.482	0.656	0.574	0.491	0.68	0.623
	RF	0.573	0.485	0.477	0.525	0.74	0.618	0.649	0.678	0.2

B. Integration of GA Based Feature Selection Method and Bagging Technique

In the next experiment, we implemented GA and bagging technique for 10 classification algorithms on 9 NASA MDP data sets. The experimental result is shown in Table III. The improved model for each classifier is highlighted width boldfaced print.

TABLE III.
AUC of 10 Classifiers on 9 Data Sets (With GA and Bagging)

Classifiers		CM1	KC1	KC3	MC2	MW1	PC1	PC2	PC3	PC4
Statistical Classifier	LR	0.753	0.795	0.691	0.761	**0.742**	0.852	0.822	0.813	**0.901**
	LDA	**0.592**	**0.627**	**0.635**	**0.64**	**0.674**	**0.637**	**0.607**	**0.635**	**0.715**
	NB	0.702	**0.79**	**0.677**	0.739	0.724	**0.799**	0.805	**0.78**	**0.861**
Nearest Neighbor	k-NN	**0.666**	**0.689**	**0.67**	**0.783**	**0.656**	**0.734**	**0.554**	**0.649**	**0.732**
	K*	**0.71**	**0.822**	0.503	**0.718**	**0.68**	**0.876**	**0.877**	**0.816**	**0.893**
Neural Network	BP	**0.744**	**0.797**	**0.707**	**0.835**	**0.689**	**0.829**	0.905	**0.799**	**0.921**
Support Vector Machine	SVM	0.667	**0.767**	0.572	**0.747**	**0.659**	0.774	0.139	0.476	**0.879**
Decision Tree	C4.5	**0.64**	**0.618**	**0.658**	**0.732**	**0.695**	**0.758**	**0.642**	**0.73**	**0.844**
	CART	**0.674**	**0.818**	**0.754**	**0.709**	**0.703**	**0.819**	**0.832**	**0.842**	**0.9**
	RF	**0.706**	**0.584**	**0.605**	0.483	0.735	**0.696**	**0.901**	**0.734**	**0.601**

Figure 4 visually shows AUC comparisons of 10 algorithms on 9 NASA MDP data sets. As shown in Table III and Figure 4, almost all classifiers that

implemented GA and bagging outperform the original method. It indicate that the integration of GA based feature selection and bagging technique is effective to improve classification performance. Support Vector Machine (SVM) techniques perform less well and no significant improvement. It indicates that feature selection and bagging are not the answer of the SVM performance problem. This result also confirmed that SVM may be underperforming as they require parameter optimization for best performance [3]. The integration between GA and bagging technique affected significantly on the performance of the class imbalance suffered classifiers, such as C4.5 and CART.

Figure 4. AUC Comparisons of 9 Data Sets Classified by 10 Classifiers (Without/With GA and Bagging)

In order to verify whether a significant difference between the proposed method (with GA and bagging) and a method without GA and bagging, the results of both methods are compared. We performed statistical t-Test (Paired Two Sample for Means) for every classifier (algorithm) pair of without/with GA and bagging on each data set. In statistical significance testing the P value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. One often "rejects the null hypothesis" when the P value is less than the predetermined significance level (α), indicating that the observed result would be highly unlikely under the null hypothesis. In this case, we set the statistical significance level (α) to be 0.05. It means that no statistically significant difference if P value > 0.05.

The result is shown in Table IV. Although there are two classifiers (LR and NB) that have no significant difference (P value > 0.05), the results have indicated that those of remaining eight classifiers (LDA, k-NN, K*, BP, SVM, C4.5, CART and RF) have significant difference (P value < 0.05). The integration between GA and bagging technique achieved higher classification accuracy for most classifiers. Therefore, we can conclude that the proposed method makes an impressive improvement in prediction performance for most classifiers.

TABLE IV.
PAIRED TWO-TAILED T-TEST OF WITHOUT/WITH GA AND BAGGING

Classifiers		*P* value of t-Test	Result
Statistical Classifier	LR	0.156	Not Sig. (α > 0.05)
	LDA	**0.00004**	**Sig. (α < 0.05)**
	NB	0.294	Not Sig. (α > 0.05)
Nearest Neighbor	k-NN	**0.00002**	**Sig. (α < 0.05)**
	K*	**0.001**	**Sig. (α < 0.05)**
Neural Network	BP	**0.008**	**Sig. (α < 0.05)**
Support Vector Machine	SVM	**0.03**	**Sig. (α < 0.05)**
Decision Tree	C4.5	**0.0002**	**Sig. (α < 0.05)**
	CART	**0.0002**	**Sig. (α < 0.05)**
	RF	**0.01**	**Sig. (α < 0.05)**

C. Integration of PSO based Feature Selection and Bagging Technique

In the next experiment, we implemented PSO and bagging technique for 10 classification algorithms on 9 NASA MDP data sets. The experimental result is shown in Table V. The improved model for each classifier is highlighted width boldfaced print.

TABLE V.
AUC OF 10 CLASSIFIERS ON 9 DATA SETS (WITH PSO AND BAGGING)

Classifiers		CM1	KC1	KC3	MC2	MW1	PC1	PC2	PC3	PC4
Statistical Classifier	LR	0.738	0.798	0.695	**0.78**	**0.751**	0.848	0.827	**0.816**	**0.897**
	LDA	0.469	**0.627**	**0.653**	**0.686**	**0.632**	**0.665**	0.571	**0.604**	**0.715**
	NB	**0.756**	**0.847**	**0.71**	0.732	**0.748**	**0.79**	**0.818**	**0.78**	**0.85**
Nearest Neighbor	k-NN	**0.632**	**0.675**	**0.578**	**0.606**	**0.648**	**0.547**	**0.594**	**0.679**	**0.738**
	K*	**0.681**	**0.792**	**0.66**	**0.725**	0.572	**0.822**	0.814	**0.809**	**0.878**
Neural Network	BP	0.7	**0.799**	**0.726**	**0.734**	**0.722**	**0.809**	0.89	**0.823**	**0.915**
Support Vector Machine	SVM	0.721	0.723	**0.67**	0.756	0.667	**0.792**	0.294	0.735	**0.903**
Decision Tree	C4.5	**0.682**	**0.606**	**0.592**	**0.648**	**0.615**	**0.732**	**0.732**	**0.78**	**0.769**
	CART	**0.611**	**0.679**	**0.787**	**0.679**	**0.682**	**0.831**	**0.794**	**0.845**	**0.912**
	RF	**0.62**	**0.604**	**0.557**	**0.533**	0.714	**0.686**	**0.899**	**0.759**	**0.558**

Figure 5 visually shows AUC comparisons of 10 algorithms on 9 NASA MDP data sets. As shown in Table V and Figure 5, almost all classifiers that implemented PSO and bagging outperform the original method. It indicate that the integration of PSO based feature selection and bagging technique is effective to improve classification performance. Support Vector Machine (SVM) techniques still perform less well and no significant improvement. The integration between PSO

and bagging technique affected significantly on the performance of the class imbalance suffered classifiers, such as C4.5 and CART.

Figure 5. AUC Comparisons of 9 Data Sets Classified by 10 Classifiers (Without/With PSO and Bagging)

In order to verify whether a significant difference between the proposed method (with PSO and bagging) and a method without PSO and bagging, the results of both methods are compared. We performed t-Test (*Paired Two Sample for Means*) for every classifier (algorithm) pair of without/with PSO and bagging on each data set. We set significance level (α) to be 0.05. The result is shown in Table VI.

TABLE VI.
PAIRED TWO-TAILED T-TEST OF WITHOUT/WITH PSO AND BAGGING

Classifiers		*P* value of t-Test	Result
Statistical Classifier	LR	0.323	Not Sig. ($P > 0.05$)
	LDA	**0.003**	**Sig. ($P < 0.05$)**
	NB	**0.007**	**Sig. ($P < 0.05$)**
Nearest Neighbor	k-NN	**0.00007**	**Sig. ($P < 0.05$)**
	K*	**0.001**	**Sig. ($P < 0.05$)**
Neural Network	BP	**0.03**	**Sig. ($P < 0.05$)**
Support Vector Machine	SVM	0.09	Not Sig. ($P > 0.05$)
Decision Tree	C4.5	**0.0002**	**Sig. ($P < 0.05$)**
	CART	**0.002**	**Sig. ($P < 0.05$)**
	RF	**0.01**	**Sig. ($P < 0.05$)**

Although there are three classifiers that have no significant difference ($P > 0.05$), the results have indicated that those of remaining eight classifiers have significant difference ($P < 0.05$). Therefore, we can conclude that the proposed method (the integration between PSO and bagging technique) makes an impressive improvement in prediction performance for most classifiers.

D. Comparison Between GA+Bagging and PSO+Bagging

Finally, In order to verify whether a significant difference between the GA+Bagging and PSO+Bagging, the results of both methods are compared. We performed t-Test (Paired Two Sample for Means) for every classifier (algorithm) pair of GA+Bagging and PSO+Bagging on each data set. We set significance level (α) to be 0.05. The result is shown in Table VII.

TABLE VII.
PAIRED TWO-TAILED T-TEST OF GA+BAGGING AND PSO+BAGGING

Classifiers		*P* value of t-Test	Result
Statistical Classifier	LR	0.25	Not Sig. ($\alpha > 0.05$)
	LDA	0.19	Not Sig. ($\alpha > 0.05$)
	NB	**0.044**	**Sig. ($\alpha < 0.05$)**
Nearest Neighbor	k-NN	0.063	Not Sig. ($\alpha > 0.05$)
	K*	0.268	Not Sig. ($\alpha > 0.05$)
Neural Network	BP	0.203	Not Sig. ($\alpha > 0.05$)
Support Vector Machine	SVM	**0.003**	**Sig. ($\alpha < 0.05$)**
Decision Tree	C4.5	0.3	Not Sig. ($\alpha > 0.05$)
	CART	0.216	Not Sig. ($\alpha > 0.05$)
	RF	0.088	Not Sig. ($\alpha > 0.05$)

Although there are two classifier that have significant difference ($P < 0.05$) (NB and SVM), the results have indicated that those of remaining eight classifiers have no significant difference ($P > 0.05$). We can conclude that there is no significant difference between PSO and GA when used as feature selection for most classifiers.

VI. CONCLUSION

A novel method that integrate metaheuristic optimization methods (genetic algorithm and particle swarm optimization) and bagging technique for software defect prediction is proposed in this paper. Genetic algorithm and particle swarm optimization are applied to deal with the noise attributes problem, and bagging technique is employed to alleviate the class imbalance problem.

We conducted a comparative study of ten classifiers which is applied to nine NASA MDP data sets with context of software defect prediction. Experimental results show us that the proposed methods, the integration between metaheuristic optimization methods and bagging, achieved higher classification accuracy. Therefore, we can conclude that the proposed method makes an impressive improvement in prediction performance for most classifiers. From the comparison result, we conclude

that there is no significant difference between particle swarm optimization and genetic algorithm when used as feature selection for most classifiers in software defect prediction.

Future research will be concerned with benchmarking and investigating the sophisticated metaheuristic optimization methods for optimizing parameter of support vector machine for software defect prediction.

REFERENCES

[1] M. McDonald, R. Musson, and R. Smith, "The practical guide to defect prevention," *Control*, pp. 260–272, 2007.

[2] C. Jones, *Applied Software Measurement: Global Analysis of Productivity and Quality*, vol. 38, no. 1. McGraw-Hill Inc., 2008, p. 662.

[3] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, "A Systematic Literature Review on Fault Prediction Performance in Software Engineering," *IEEE Transactions on Software Engineering*, vol. 38, no. 6, pp. 1276–1304, Nov. 2012.

[4] C. Catal, "Software fault prediction: A literature review and current trends," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4626–4636, Apr. 2011.

[5] T. Menzies, Z. Milton, B. Turhan, B. Cukic, Y. Jiang, and A. Bener, "Defect prediction from static code features: current results, limitations, new approaches," *Automated Software Engineering*, vol. 17, no. 4, pp. 375–407, May 2010.

[6] F. Shull, V. Basili, B. Boehm, A. W. Brown, P. Costa, M. Lindvall, D. Port, I. Rus, R. Tesoriero, and M. Zelkowitz, "What we have learned about fighting defects," in *Proceedings Eighth IEEE Symposium on Software Metrics 2002*, 2002, pp. 249–258.

[7] Q. Song, Z. Jia, M. Shepperd, S. Ying, and J. Liu, "A General Software Defect-Proneness Prediction Framework," *IEEE Transactions on Software Engineering*, vol. 37, no. 3, pp. 356–370, May 2011.

[8] N. B. Ebrahimi, "On the statistical analysis of the number of errors remaining in a software design document after inspection," *Software Engineering IEEE Transactions on*, vol. 23, no. 8, pp. 529–532, 1997.

[9] O. Kutlubay, B. Turhan, and A. B. Bener, "A Two-Step Model for Defect Density Estimation," in *Software Engineering and Advanced Applications 2007 33rd EUROMICRO Conference on*, 2007, pp. 322–332.

[10] L. C. Briand, K. El Emam, B. G. Freimut, and O. Laitenberger, "A comprehensive evaluation of capture-recapture models for estimating software defect content," *IEEE Transactions on Software Engineering*, vol. 26, no. 6, pp. 518–540, 2000.

[11] K. El Emam and O. Laitenberger, "Evaluating capture-recapture models with two inspectors," *IEEE Transactions on Software Engineering*, vol. 27, no. 9, pp. 851–864, 2001.

[12] G. Rücker, V. Reiser, E. Motschall, H. Binder, J. J. Meerpohl, G. Antes, and M. Schumacher, "Boosting qualifies capture-recapture methods for estimating the comprehensiveness of literature searches for systematic reviews.," *Journal of Clinical Epidemiology*, vol. 64, no. 12, pp. 1364–72, 2011.

[13] P. E. R. Runeson and C. Wohlin, "An Experimental Evaluation of an Experience-Based Capture-Recapture Method in Software Code Inspections," *Empirical Software Engineering*, vol. 3, no. 4, pp. 381–406, 1998.

[14] N. E. Fenton and M. Neil, "A critique of software defect prediction models," *IEEE Transactions on Software Engineering*, vol. 25, no. 5, pp. 675–689, 1999.

[15] M. Shepperd, M. Cartwright, and C. Mair, "Software defect association mining and defect correction effort prediction," *IEEE Transactions on Software Engineering*, vol. 32, no. 2, pp. 69–82, Feb. 2006.

[16] R. Karthik and N. Manikandan, "Defect association and complexity prediction by mining association and clustering rules," *2010 2nd International Conference on Computer Engineering and Technology*, pp. V7–569–V7–573, 2010.

[17] C.-P. Chang, C.-P. Chu, and Y.-F. Yeh, "Integrating in-process software defect prediction with association mining to discover defect pattern," *Information and Software Technology*, vol. 51, no. 2, pp. 375–384, Feb. 2009.

[18] A. A. Porter and W. Richard, "Empirically Guided Software Development Using Metric-Based Classification Trees," *IEEE Software*, pp. 46–54, 1990.

[19] L. Zhan and M. Reformat, "A practical method for the software fault-prediction," in *IEEE International Conference on Information Reuse and Integration*, 2007, pp. 659–666.

[20] T. Menzies, J. Greenwald, and A. Frank, "Data Mining Static Code Attributes to Learn Defect Predictors," *IEEE Transactions on Software Engineering*, vol. 33, no. 1, pp. 2–13, Jan. 2007.

[21] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings," *IEEE Transactions on Software Engineering*, vol. 34, no. 4, pp. 485–496, Jul. 2008.

[22] N. Gayatri, S. Nickolas, A. Reddy, S. Reddy, and A. V. Nickolas, "Feature Selection Using Decision Tree Induction in Class level Metrics Dataset for Software Defect Predictions," *Proceedings of the World Congress on Engineering and Computer Science*, vol. I, pp. 124–129, 2010.

[23] G. Denaro, "Estimating software fault-proneness for tuning testing activities," in *Proceedings of the 22nd International Conference on Software engineering - ICSE '00*, 2000, pp. 704–706.

[24] T. M. Khoshgoftaar and K. Gao, "Feature Selection with Imbalanced Data for Software Defect Prediction," *2009 International Conference on Machine Learning and Applications*, pp. 235–240, Dec. 2009.

[25] B.-J. Park, S.-K. Oh, and W. Pedrycz, "The design of polynomial function-based neural network predictors for detection of software defects," *Information Sciences*, Jan. 2011.

[26] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 3, pp. 552–568, May 2011.

[27] A. Tosun, A. Bener, B. Turhan, and T. Menzies, "Practical considerations in deploying statistical methods for defect prediction: A case study within the Turkish telecommunications industry," *Information and Software Technology*, vol. 52, no. 11, pp. 1242–1257, Nov. 2010.

[28] T. M. Khoshgoftaar, Y. Xiao, and K. Gao, "Software quality assessment using a multi-strategy classifier," *Information Sciences*, Dec. 2010.

[29] Y. Liu, X. Yu, J. X. Huang, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Information Processing & Management*, vol. 47, no. 4, pp. 617–631, Jul. 2011.

[30] T. Wang, W. Li, H. Shi, and Z. Liu, "Software Defect Prediction Based on Classifiers Ensemble," *Journal of Information & Computational Science 8:*, vol. 16, no. December, pp. 4241–4254, 2011.
[31] S. Kim, H. Zhang, R. Wu, and L. Gong, "Dealing with noise in defect prediction," *Proceeding of the 33rd International Conference on Software engineering - ICSE '11*, pp. 481–490, 2011.
[32] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson, "Reflections on the NASA MDP data sets," *IET Software*, vol. 6, no. 6, p. 549, 2012.
[33] S. C. Yusta, "Different metaheuristic strategies to solve the feature selection problem," *Pattern Recognition Letters*, vol. 30, no. 5, pp. 525–534, Apr. 2009.
[34] M. M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3747–3763, Feb. 2012.
[35] H. Wang, T. M. Khoshgoftaar, and A. Napolitano, "Software measurement data reduction using ensemble techniques," *Neurocomputing*, vol. 92, pp. 124–132, Sep. 2012.
[36] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Improving Software-Quality Predictions With Data Sampling and Boosting," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 39, no. 6, pp. 1283–1294, Nov. 2009.
[37] K. Gao, T. M. Khoshgoftaar, H. Wang, and N. Seliya, "Choosing software metrics for defect prediction: an investigation on feature selection techniques," *Software: Practice and Experience*, vol. 41, no. 5, pp. 579–606, Apr. 2011.
[38] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[39] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings IEEE World Congress on Computational Intelligence Cat No98TH8360*, 1998, vol. 189, no. 2, pp. 69–73.
[40] C. Catal and B. Diri, "Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem," *Information Sciences*, vol. 179, no. 8, pp. 1040–1058, Mar. 2009.
[41] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. Elsevier Inc., 2011.
[42] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*, vol. 12. Springer-Verlag Berlin Heidelberg, 2011.

Romi Satria Wahono. Received B.Eng and M.Eng degrees in Computer Science in 1999 and 2001, respectively from Saitama University, Japan. Currently working towards a PhD with Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. He is a lecturer at the Graduate School of Computer Science, Dian Nuswantoro University, Indonesia. He is also a founder and chief executive officer of Brainmatics, Inc., a software development company in Indonesia. His current research interests include software engineering and machine learning. Professional member of the ACM and IEEE Computer Society.

Nanna Suryana. Received his B.Sc. in Soil and Water Eng. (Bandung, Indonesia), M.Sc. in Comp. Assisted for Geoinformatics and Earth Science, (Enschede, Holland), Ph.D. in Geographic Information System (GIS) (Wageningen, Holland). He is currently holding a position of Director of International Office and professor at Faculty of Information Technology and Communication (FTMK) of Universiti Teknikal Malaysia Melaka (UTEM). His current research interest is in field of GIS and Data Mining.

Sabrina Ahmad. Received BIT (Hons) from Universiti Utara Malaysia and MSc. in real time software engineering from Universiti Teknologi Malaysia. She obtained Ph.D in Computer Science from The University of Western Australia. Her specialization is in requirements engineering and focusing on improving software quality. She is currently a Senior Lecturer at Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. Her research interests include software engineering, software requirements, quality metrics and process model. Certified Professional Requirements Engineering (IREB CPRE -FL) and Certified Information Technology Architect (IASA CITA-F).