Evaluating Contractor Financial Status Using a Hybrid Fuzzy Instance Based Classifier: Case Study in the Construction Industry

Min-Yuan Cheng and Nhat-Duc Hoang

***Abstract*—Construction firms are vulnerable to bankruptcy due to the complex nature of the industry, high competitions, the high risk involved, and considerable economic fluctuations. Thus, evaluating financial status and predicting business failures of construction companies are crucial for owners, general contractors, investors, banks, insurance firms, and creditors. The prediction results can be used to select qualified contractors capable of accomplishing the projects. In this study, a hybrid fuzzy instance-based classifier for contractor default prediction (FICDP) is proposed. The new approach is constructed by incorporating the fuzzy *K*-nearest neighbor classifier (FKNC), the synthetic minority over-sampling technique (SMOTE), and the firefly algorithm (FA). In this hybrid paradigm, the FKNC is utilized to classify the contractors into two groups ("default" and "nondefault") based on their past financial performances. Since the "nondefault" samples dominate the historical database, the SMOTE algorithm is employed to create synthetic samples of the minority class and therefore alleviates the between-class imbalance problem. Moreover, the FA is employed to determine an appropriate set of model parameters. Experimental results have shown that the proposed FICDP can outperform other benchmark methods.**

***Index Terms*—Default prediction, fuzzy instance based classifier, swarm intelligence, synthetic minority over-sampling technique.**

I. Background

EVALUATING the financial status of construction contractors is a very crucial task in construction project management. The reason is that project owners or managers want to avoid awarding contracts to the contractors with high failure tendency [1]. Furthermore, the increasing scale and technical complexity of construction projects often leads to the collaboration of many contractors to accomplish the projects [2]. Thus, a successful construction project is highly dependent on the cooperation of prime contractor and subcontractors, if one contractor or subcontractor defaults, the others will also be affected [1].

As a consequence, default prediction has become an attractive research domain in the academic community [3]. Mason and Harris [4] and Abidali and Harris [5] attempted to predict company failure in the construction industry with the multiple discriminant analysis (MDA). Hall [6] carried out a survey on small construction firms in the United Kingdom and constructed a failure prediction model based on logistic regression (LR). The LR approach has also been applied to forecast the trend of company decline [7] and contractor default probability [8].

Recently, artificial intelligence (AI) approaches have been successfully employed for financial healthiness appraisal [1], [9]. Prediction models based on AI techniques have been proved to be a feasible and reliable alternative for the problem at hand. Using AI techniques, this problem can be modeled as a classification task in which prediction outputs are either "default" or "nondefault." By learning the data recorded in the past, the AI models are capable of forecasting results based on the new input patterns [10].

Al-Sobiei *et al*. [9], [11] utilized the artificial neural networks (ANN) to estimate the risk of contractor default. Lam *et al*. [12] and Horta and Camanho [3] employed the support vector machine (SVM) techniques to develop decision support systems for assessing contractor's financial status. Tserng *et al*. [1] established models to predict the bankruptcy status of construction firms based on the LR and the SVM; the research found that the SVM methods outperform the LR. Lam and Yu [13] proposed a prediction model based on the multiple kernel learning for contractor prequalification.

The ANN approach a major disadvantage is that its training process is achieved through a gradient descent algorithm on the error space, which can be very complex and may contain many local minima [14]. Thus, the ANN training process is likely to be trapped in some local minima and this can deteriorate the financial status prediction performance. Additionally, the SVM training process entails solving a quadratic programming problem subjected to inequality constraint. This means that the SVM's training process for large data sets requires expensive computational cost. More importantly, the black box nature of the ANN and SVM approaches makes them difficult for practical project managers or owners to comprehend how the models predict the contractor default status.

Different from the SVM method, the fuzzy *K*-nearest neighbor classifier (FKNC) [15] belongs to the class of instance-based learning. This algorithm employs the collected historical cases to establish its memory. A FKNC utilizes the information obtained from the *K*-nearest neighbors of a sample vector and assigns class memberships to it. The vector's membership

Manuscript received August 18, 2014; revised November 9, 2014; accepted December 16, 2014. Date of publication February 5, 2015; date of current version April 15, 2015. Review of this manuscript was arranged by Department Editor B. Jiang.

M.-Y. Cheng is with the Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taipei 10608, Taiwan (e-mail: myc@mail.ntust.edu.tw).

N.-D. Hoang is with the Institute of Research and Development, Faculty of Civil Engineering, Duy Tan University, Danang 550000, Vietnam (e-mail: hoangnhatduc@dtu.edu.vn).

Digital Object Identifier 10.1109/TEM.2014.2384513

values supply a level of assurance to accompany the classification outcome.

Moreover, the FKNC also assigns fuzzy memberships as a function of the vector's distance from its *K*-nearest neighbors and those neighbors' memberships in the possible classes [16]. This approach is simple to implement and its classification outcomes can be much easier to interpret. Moreover, the competitiveness of the FKNC approach compared with the ANN, SVM, LR, and probabilistic neural network algorithms has been demonstrated in various applications [15], [17]–[19]. Nevertheless, none of previous works has evaluated the capability of this technique in contractor default assessment. Therefore, this research is an attempt to fill this gap.

This paper extends previous literature by proposing an innovative construction contractor default prediction model that employs the FKNC as the instance-based learning method. It is noted that the implementation of the FKNC requires a proper setting of two tuning parameters: the neighboring size (*k*) and the fuzzy strength (*m*). Furthermore, this parameter selection process can be modeled as an optimization problem. Metaheuristic approaches, such as the genetic algorithm [20]–[22], particle swarm optimization [23], [24], differential evolution [17], [25], [26] algorithms have been shown to be feasible to tackle the optimization problem at hand.

Among metaheuristic methods, the firefly algorithm (FA), recently developed by Yang [27], is a fast and effective metaheuristic for solving global optimization in continuous space. Numerical experiments in previous researches have demonstrated the superior performance of the FA over other optimization methods [28]–[31]. Nonetheless, few research works have investigated the capability of this algorithm in dealing with the parameter selection problem. Thus, this study proposes to hybridize the FKNC with the FA [27] to automatically search for appropriate control parameters of the prediction model.

Furthermore, in the historical database, the number of "non-default" construction firms occupies more than 98% of the collected samples. In order to balance the class distribution, undersampling and oversampling methods are often employed [32], [33]. The undersampling methods aims at creating a subset of the original dataset by eliminating instances, which are usually majority class instances; the major drawback of these techniques is that they may discard potentially useful data which can be crucial for the learning process [34].

To deal with this drawback, sophisticated oversampling methods have been put forward. Among them, the synthetic minority over-sampling technique (SMOTE) has been proved to be very effective and has been utilized successfully by researchers in various fields [35]–[37]. In brief, the main idea of this technique is to create new minority class examples by interpolating several minority class instances that lie together for oversampling the training set [34], [38]. Hence, the proposed fuzzy instance-based classifier for contractor default prediction (FICDP) integrates the SMOTE to overcome the between-class imbalance problem. The rest of this paper is organized as follows: the second section of this paper reviews the research methodology. The third section presents the historical dataset of financial records of construction firms. The next section provides detailed description of proposed model. The model application is reported in the fifth section. Some conclusions of this study are stated in the final part.

II. Research Methodology

A. Fuzzy K-Nearest Neighbor Classifier

The FKNC algorithm is an instance-based classifier that incorporates the fuzzy set theory into the classification process [15]. In the FKNC, the fuzzy memberships of samples are assigned to different classes. The class which possesses the maximum membership degree can be chosen as the winner. The first step of the FKNC is to calculate the fuzzy partition matrix $U = [u_{ij}]$ from the memory, which stores a set of *n* training sample vectors $[x_1, \ldots, x_n]$. Herein, we denote *j* as the vector index $(j = 1, 2, \ldots, n)$, where *n* is the number of training samples, and, the variable *i* represents the class index $(i = 1, 2, \ldots, C)$, where *C* is the number of classes. For each training case *x*, we identify its *K*-nearest neighbors by calculating Euclidean distances. The membership degree of the sample vector x_j in the class *i* is given as follows:

$$u_{ij}(x) = u_i(x_j) = \begin{cases} 0.51 + (n_i/K) \times 0.49, \text{ if } c(x_j) = i \\ (n_i/K) \times 0.49, \text{ if } c(x_j) \neq i \end{cases} \tag{1}$$

where n_i is the number of neighbors found which belong to the class *i* and $c(x_j)$ represents the class label of the sample vector x_j. It is obvious that u_{ij} is an element of the $C \times n$ matrix *U*. Moreover, it is also worth noticing that the purpose of Eq. (1) is to assign higher fuzzy membership grades to the training samples that stay away from the decision boundary and lower fuzzy memberships grade to the patterns that lie in the vicinity of the decision boundary [15]. It is because the information supplied by the samples in the region close to the decision surface is more uncertain than that provided by other samples.

Since u_{ij} is a fuzzy membership grade of the sample x_j in the class *i*, u_{ij} must satisfy the following properties:

$$u_{ij} \in [0, 1] \tag{2}$$

$$\sum_{i=1}^{C} u_{ij} = 1 \tag{3}$$

$$0 < \sum_{j=1}^{n} u_{ij} < n. \tag{4}$$

The second step of the FKNC is to assign fuzzy memberships of the unknown sample *x* to different classes according to the following equation:

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij} \left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)}{\sum_{j=1}^{K} \left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)} \tag{5}$$

where $i = 1, 2, \ldots, C$, and $j = 1, 2, \ldots, K$. Herein, *j* represents the *j*th sample vector among the *K*-nearest neighbors

Begin FA
Define objective function $f(x)$, where $x=(x_1,...,x_d)$
Generate an initial population of fireflies
Formulate the light intensity I
Define the absorption coefficient γ
While ($t <$ Max_Generation)
For $i = 1$ to n (all n fireflies)
  For j=1 to n (all n fireflies)
    If ($I_j > I_i$), move firefly i towards firefly j
    End if
    Evaluate new solutions and update light intensity;
  End for j
End for i
Rank the fireflies and find the current best
End while;
End FA

Fig. 1. FA algorithm.

of x, C is the number of classes, and K denotes the neighboring size. The fuzzy strength m is used to determine how heavily the distance is weighted when computing each neighbor's contribution to the membership value. $||x - x_j||$ represents the distance between x and its jth nearest neighbor x_j. In this study, Euclidean metric is used as the distance measurement. u_{ij} denotes the membership degree of the sample vector x_j in the class i and is computed in the first step of the algorithm (refer to Eq. (1)).

B. Firefly Algorithm

In order to commence the training process of the FKNC, two tuning parameter (K and m) are required to be determined. A proper setting of these tuning parameters is necessary to achieve a desirable performance of the prediction model [14]. Thus, in this study, we utilize the FA as a means for tuning the FKNC's free parameters. The description of the FA algorithm is provided in the following section of this paper.

The flashing lights of fireflies are an amazing sight in the summer sky in tropical and temperate regions. The pattern of flashes is often unique for a particular species. In essence, each firefly is attracted to brighter ones as it randomly explores while searching for prey. Based on that phenomenon in nature, the FA is formulated as a global optimization method. The FA is an advanced swarm intelligence that can locate the optimum effectively [27], [39]. Superior performance of the FA over other metaheuristic algorithms has been demonstrated in previous research works [29], [40], [41].

The FA utilizes the following rules.

1) All fireflies are unisex, so each firefly is attracted to other fireflies regardless of their sex.
2) The attractiveness of a firefly is proportional to its brightness and decreases as the distance increases. A firefly moves randomly if no other firefly is brighter.
3) The brightness of a firefly is affected or determined by the landscape of the objective function.

The FA pseudocode is illustrated in Fig. 1.

The brightness of an individual firefly can be defined similarly to the fitness value in the genetic algorithm. The light intensity $I(r)$ varies according to the following equation:

$$I(r) = I_o \exp(-\gamma r^2) \tag{6}$$

where I_o denotes the light intensity of the source, γ is the light absorption coefficient, and r represents the distance from the source.

As the attractiveness of a firefly is proportional to the light intensity seen by adjacent fireflies, the attractiveness β of a firefly is defined as

$$\beta = \beta_o \exp(-\gamma r^2). \tag{7}$$

In a D-dimensional space, the distance between any two fireflies i at x_{i} and j at x_{j}, is the calculated as follows:

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{D} (x_{i,k} - x_{j,k})^2}. \tag{8}$$

Since a specific firefly x_i is attracted to the brighter one x_j, the movement of the ith firefly can be expressed as

$$x_i = x_i + \beta_o \exp(-\gamma r_{ij}^2)(x_i - x_j) + \alpha(\omega - 0.5) \tag{9}$$

where γ is the light absorption coefficient, γ varies from 0.1 to 10, β_0 represents the attractiveness at $r_{ij} = 0$, α denotes a tradeoff constant to determine the random behavior of movement, ω represents a random number drawn from the Gaussian distribution.

C. SMOTE for Dealing With the Imbalanced Classification Problem

To establish classification model with datasets which suffer from imbalanced class distributions is an important problem in the field of data mining [34]. This problem occurs when the number of instances representing the class of interest is much

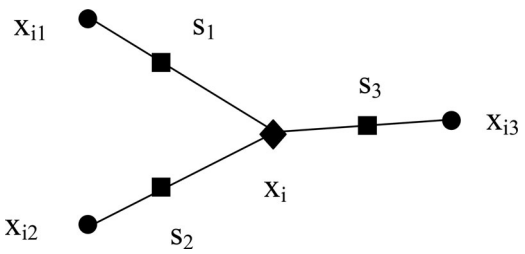

Fig. 2. Illustration of the SMOTE algorithm.

TABLE I
FINANCIAL VARIABLES

	Liquidity		Leverage
X_1	Current ratio	X_5	Total liabilities to net worth
X_2	Quick ratio	X_6	Retained earnings to sales
X_3	Net working capital to total assets	X_7	Debt ratio
X_4	Current assets to net assets	X_8	Times interest earned
	Activity		Profitability
X_9	Revenues to net working capital	X_{17}	Return on assets
X_{10}	Accounts receivable turnover	X_{18}	Return on equity
X_{11}	Accounts payable turnover	X_{19}	Return on sales
X_{12}	Sales to net worth	X_{20}	Profits to net working capital
X_{13}	Quality of inventory		
X_{14}	Fixed assets to net worth		
X_{15}	Turnover of total assets		
X_{16}	Revenues to fixed assets		

lower than the ones of the other classes. Many real-world classification problems (e.g., disease diagnosis, fraudulent telephone calls, oil spills, structure collapses, etc.) are imbalanced in which the minority class is usually the one that has the highest interest from a learning point of view and it also implies a great cost when it is not well classified [38], [42], [43]. The minority class usually represents the most important concept to be learned, and it is difficult to identify it since it might be associated with exceptional and significant cases or because the data acquisition of these examples is costly [34], [44].

The SMOTE [38] is a renowned method to cope with the between-class imbalance problem. This technique creates new samples of the minority class by interpolating several minority class instances that lie together for oversampling the training set. Hence, the minority class is oversampled by introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of oversampling required ($N\%$), neighbors from the k-nearest neighbors are randomly chosen. The SMOTE algorithm is demonstrated in Fig. 2, where X_i is the selected point, X_{i1}, Xx_{i2}, and X_{i3} are selected nearest neighbors and S_1, S_2, and S_3 are the synthetic data points created by the randomized interpolation.

D. Evaluating Classification Performance for Imbalanced Classification Problems

In a common classification task, the accuracy rate, quantified by the ratio of correctly predicted cases over the total number of cases, can be used to measure the classifier performance [25], [34]. However, in the context of imbalanced classification problems, the accuracy rate is not a proper indicator of a good classifier, since it does not distinguish between the number of correctly classified examples of different classes [34]. Hence, the accuracy rate may lead to delusive conclusions, in other words, a very high accuracy rate, e.g., 100%, may indicate a very poor classifier.

When facing imbalanced datasets, the evaluation of the classifiers' performance must be carried out using specific metrics in order to take into account the imbalanced class distribution. The following four metrics can be used to measure the classification performance [34]: true positive rate (the percentage of positive instances correctly classified), true negative rate (the percentage of negative instances correctly classified), false positive rate (the percentage of negative instances misclassified), and false negative rate (the percentage of positive instances misclassified).

This four metrics can be summarized in a confusion matrix [45]. Moreover, a well-known approach to incorporate these four measures and to produce an evaluation criterion is to employ the receiver operating characteristic (ROC) curve [46]. In addition, the area under the ROC curve, denoted as AUC, provides a single measure of a classifier's performance for evaluating which model is better on average [1], [47]. It is worth noticing that a higher AUC value indicates a better predictive performance. Generally, a classifier with perfect predictive ability has an AUC of 1, meanwhile, a poor classifier with random predictions has an AUC of 0.5. Moreover, an AUC of the range (0.7, 0.8) indicates an acceptable classification performance. If $0.8 \leq \text{AUC} \leq 0.9$, an excellent classification performance is attained. If $\text{AUC} \geq 0.9$, the classifier has attained an outstanding performance.

III. HISTORICAL DATA OF CONSTRUCTION COMPANY FINANCIAL RECORDS

This research utilizes a database of construction firms collected from the Wharton Research Data Services 2011. The selected companies were categorized by construction types defined in the standard industrial classification (SIC) numerical codes ranging from 1500–1799. Additionally, it is noted that financial data is taken at fiscal year-ends. The selected contractors include three construction categories: SIC code 1500–1599 (for building construction, general contractors, and operative builders), SIC code 1600–1699 (for heavy construction other than building construction contractors), and SIC code 1700–1799 (for construction special trade contractors).

The dataset consists of 76 construction companies within which 63 companies are nondefaulted and 13 companies are defaulted. The financial data of each firm are summarized and recorded yearly. All available financial firm records in the database are considered in this research. The time period of available financial data represented ranges from 1970–2011. In total, 958 firm-year observations have been collected. Inherent from previous works, in this study, the term "default" is defined by the CRSP delisting code of 400 and 550 to 585; the defaulted firms are those delisted because of bankruptcy, liquidation, or poor performance [48], [49]. In this research, the financial data

TABLE II
DEFINITIONS OF FINANCIAL VARIABLES

	Variables	Definition
X_1	Current Ratio	$\frac{\text{Current Assets}}{\text{Current Liabilities}}$
X_2	Quick Ratio	$\frac{\text{Current Assets}-\text{Inventories}}{\text{Current Liabilities}}$
X_3	Net Working Capital to Total Assets	$\frac{\text{Current Assets}-\text{Current Liabilities}}{\text{Total Assets}}$
X_4	Current Asset to Net Assets	$\frac{\text{Current Assets}}{\text{Total Assets}-\text{Current Liabilities}}$
X_5	Total Liabilities to Net Worth	$\frac{\text{Total Liabilities}}{\text{Net Worth}}$
X_6	Retained Earnings to Sales	$\frac{\text{Retained Earnings}}{\text{Net Sales}}$
X_7	Debt Ratio	$\frac{\text{Total Liabilities}}{\text{Total Assets}}$
X_8	Times Interest Earned	$\frac{\text{Earnings Before Interest Expense}}{\text{Interest Expense}}$
X_9	Revenues to Net Working Capital	$\frac{\text{Net Sales}}{\text{Average Current Assets}-\text{Average Current Liabilities}}$
X_{10}	Accounts Receivable Turnover	$\frac{\text{Net Sales}}{\text{Average Receivables}}$
X_{11}	Accounts Payable Turnover	$\frac{\text{Net Sales}}{\text{Average Payables}}$
X_{12}	Sales to Net Worth	$\frac{\text{Net Sales}}{\text{Average Net Worth}}$
X_{13}	Quality of Inventory	$\frac{\text{Cost of Sales}}{\text{Average Inventories}}$
X_{14}	Fixed Assets to Net Worth	$\frac{\text{Fixed Assets}}{\text{Net Worth}}$
X_{15}	Turnover of Total Assets	$\frac{\text{Net Sales}}{\text{Average Total Assets}}$
X_{16}	Revenues to Fixed Assets	$\frac{\text{Net Sales}}{\text{Average Fixed Assets}}$
X_{17}	Return of Assets (ROA)	$\frac{\text{Net Profit After Interest and Taxes}+\text{Interest Expense}}{\text{Total Assets}}$
X_{18}	Return of Equity (ROE)	$\frac{\text{Net Profit After Interest and Taxes}}{\text{Net Worth}}$
X_{19}	Return of Sales (ROS)	$\frac{\text{Net Profit After Interest and Taxes}}{\text{Net Sales}}$
X_{20}	Profits to Net Working Capital	$\frac{\text{Net Profit After Interest and Taxes}}{\text{Current Assets}-\text{Current Liabilities}}$

in the year before the companies are delisted are considered as default samples.

It is noted that financial items in database are listed in financial statements including balance sheets, income statements, and statements of cash flows. The financial ratios are then calculated from the financial statement. Financial ratios can be employed to describe financial characteristics and to evaluate performances of construction companies. Additionally, the selected financial items can serve as indicators for deriving financial ratios which later become input attributes of a default prediction model.

In this research, a set of 20 financial ratios are selected as input attributes for default prediction. The financial ratios and their financial items can be seen in Table I and the detailed equations used for computing these ratios are shown in Table II. The statistical descriptions of the data are illustrated in Table III. The reason to use these 20 financial ratios as input variables is that they have been used in previous studies involving contractor default status prediction and these variables encompass a broad cross-section of accounting ratios, which describe a contractor's liquidity, leverage, activity, and profitability [1], [50]–[53].

TABLE III
STATISTICAL SUMMARY OF FINANCIAL VARIABLES

	Financial Variable	Min	Max	Mean	St. dev.
X_1	Current Ratio	0.0242	23.371	1.9491	1.4594
X_2	Quick Ratio	0.0076	9.1006	1.3771	1.0491
X_3	Net Working Capital to Total Assets	−1.974	0.8763	0.2205	0.2128
X_4	Current Assets to Net Assets	−137.9	23.136	0.9509	5.0207
X_5	Total Liabilities to Net Worth	−97.03	1225.6	4.196	43.645
X_6	Retained Earnings to Sales	−62.69	1.7276	−0.075	2.1326
X_7	Debt Ratio	0.0597	7.9549	0.6071	0.3253
X_8	Times Interest Earned	−541	8873	37.771	420.65
X_9	Revenues to Net Working Capital	−1617	1607.8	10.658	86.364
X_{10}	Accounts Receivable Turnover	0.0214	383.95	12.006	26.419
X_{11}	Accounts Payable Turnover	0.0431	215.23	16.949	14.592
X_{12}	Sales to Net Worth	−1514	176.17	3.6331	51.174
X_{13}	Quality of Inventory	0.0633	539.97	23.953	41.827
X_{14}	Fixed Assets to Net Worth	−13.86	583.52	1.4213	18.964
X_{15}	Turnover of Total Assets	0.0156	6.0712	1.7491	0.9425
X_{16}	Revenues to Fixed Assets	0.0752	374.76	18.682	32.377
X_{17}	Return on Assets (ROA)	−2.397	0.4329	0.0316	0.154
X_{18}	Return on Equity (ROE)	−78.4	15.892	−0.17	3.4096
X_{19}	Return on Sales (ROs)	−4.968	0.4474	−0.014	0.2618
X_{20}	Profits to Net Working Capital	−17.33	632.53	1.2276	21.55

IV. FUZZY INSTANCE-BASED CLASSIFIER FOR CONTRACTOR DEFAULT PREDICTION

This section describes the proposed contractor default prediction model, named as FICDP, in detail. The model (see Fig. 3) is established by a hybridization of the FKNC, the FA optimization algorithm, and the SMOTE. The FICDP employs the FKNC as the instance-based learning technique for carrying out classification tasks. In addition, the model incorporates the FA algorithm for automatically identifying the optimal values of tuning parameters. It is worth noticing that the construction of the FICDP is dependent on two tuning parameters, namely the neighboring size (k), and the fuzzy strength (m). Meanwhile, the SMOTE algorithm is utilized to tackle the imbalanced classification problem.

1) *Input Data*: The dataset contains 958 firm-year observations within which 945 samples are nondefaulted and 13 samples are defaulted. The input data provides the

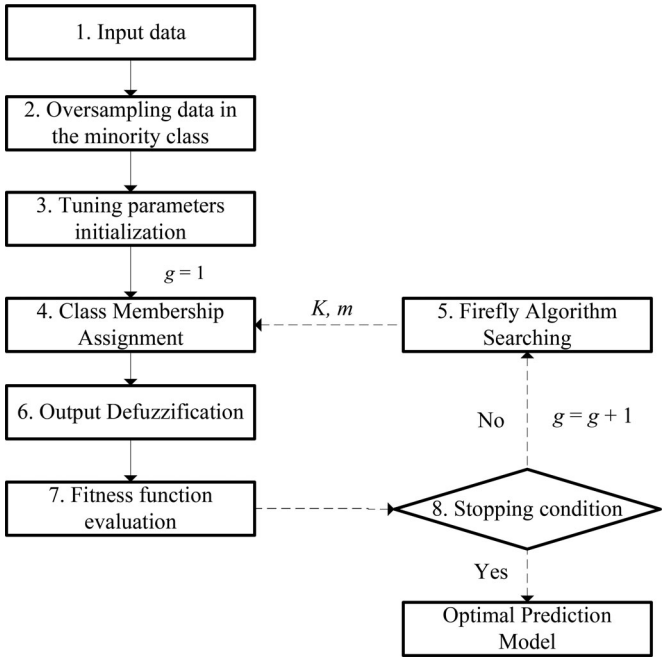

Fig. 3. Fuzzy instance-based classifier for contractor default prediction (FICDP).

financial variables that are employed to predict the default status. The data are real values and are normalized into a range of (0, 1). This transformation aims to help prevent the situation in which attributes with greater numeric magnitudes dominate those with smaller magnitudes. It is worth noticing that the ratio of nondefault and default samples is 945:13; this is widely known as the between-class imbalance problem in which the default samples belong to the minority class.

2) *Oversampling of Data in the Minority Class:* To overcome the between-class imbalance problem, the oversampling technique SMOTE is employed to obtain a balanced dataset. The number of nearest neighbors (*k*), used in the SMOTE algorithm, is 13 which are equal to the number of default samples. Moreover, the oversampling amount (*N*) is chosen so that it helps to attain the most balanced default/nondefault ratio. The oversampling amount is selected as 7300%. After being processed, the number of samples in the minority class becomes $12 \times 73 = 949$. Thus, nondefault/default ratio becomes 945:949. The whole data is randomly divided into the training and testing sets. The training set is used to construct the prediction model; meanwhile, the testing set is utilized to verify the model performance.
3) *Tuning Parameter Initialization*: The aforementioned tuning parameters of the model are randomly generated within the range of lower and upper boundaries. In this study, the lower and upper boundaries of the neighboring size (*k*) are 1 and 30, respectively. Meanwhile, these two values of the fuzzy strength (*m*) are 1.0001 and 10. Moreover, the equation used for generating the model tuning parameters can be shown as follows:

$$X_{i,0} = LB + \text{rand}[0,1] \times (UB - LB) \tag{10}$$

where $X_{i,0}$ is the tuning parameter *i* at the first generation. rand[0,1] denotes a uniformly distributed random number between 0 and 1. *LB* and *UB* are two vectors of lower bound and upper bound for any parameter.
4) *Class Membership Assignment*: In this step, the FKNC algorithm is deployed to assign fuzzy memberships of an input vector to different classes. This step requires two parameters (the neighboring size and the fuzzy strength) that are acquired from the FA component. It is noted that the problem at hand is a two-class classification task with two labels: "default" and "nondefault." Thus, for each input pattern *x*, there are two outputs, $u_1(x)$ and $u_2(x)$, representing membership degrees of *x* in the two classes.
5) *Firefly Algorithm Searching*: The FA optimization technique is applied to automatically explore the various combinations of the tuning parameters (*k* and *m*). At each generation, the optimizer carries out its searching process to guide the population of fireflies to the optimal solution. By evaluating the fitness of each firefly, the algorithm discards inferior combinations of *m* and *k*, and permits robust combinations of these parameters to be passed on the next generations.
6) *Output Defuzzification*: Because the FKNC yields fuzzy memberships of an input pattern in the two classes ($u_1(x)$ and $u_2(x)$), a step of defuzzification is employed to convert fuzzy outputs to crisp outputs ($Y(x)$) as follows:

$$Y(x) = \arg\max_{i=1}^{2}(u_i(x)). \tag{11}$$

7) *Fitness Evaluation*: In this step, the training dataset is divided into five mutually exclusive subsets. In each run, one subset is used as a validating set; meanwhile, the other subsets are used for constructing the model memory. In order to determine the optimal tuning parameters of the FKNC, the following objective function is used:

$$F_{\text{fitness}} = \frac{1}{\sum_{k=1}^{5} \text{AUC}_k} \tag{12}$$

where AUC_k denotes the AUC values obtained from the prediction results of the validating set in the *k*th run.
8) *Stopping Condition*: The optimization process of the FA algorithm terminates when the maximum number of generation is achieved. If the stopping condition is not met, the FA will continue its searching progress. When the program terminates, the optimal set of tuning parameters has been successfully identified. The FICDP is ready to predict new input patterns.

V. Experimental Results and Comparison

In the experiment, the whole database is randomly divided into two sets: the training and testing sets. The training set is used to establish the prediction model; the model performance

TABLE IV
FICDP'S PREDICTION RESULTS

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	Y_A	Y_P
0.044	0.098	0.689	0.894	0.080	0.968	0.105	0.057	0.489	0.009	0.030	0.534	0.012	0.024	0.443	0.061	0.833	0.823	0.908	0.025	1	1
0.087	0.216	0.778	0.860	0.074	0.975	0.053	0.058	0.503	0.011	0.077	0.898	0.111	0.024	0.257	0.022	0.878	0.833	0.926	0.027	0	0
0.057	0.142	0.748	0.863	0.074	0.975	0.060	0.058	0.507	0.026	0.077	0.900	0.345	0.024	0.578	0.027	0.861	0.832	0.919	0.027	0	0
0.043	0.135	0.683	0.859	0.074	0.964	0.061	0.056	0.501	0.018	0.070	0.898	0.058	0.024	0.224	0.030	0.710	0.822	0.859	0.022	1	1
0.081	0.224	0.830	0.890	0.075	0.950	0.075	0.056	0.493	0.012	0.032	0.522	0.076	0.023	0.306	0.263	0.565	0.810	0.809	0.021	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.073	0.157	0.734	0.858	0.074	0.978	0.060	0.058	0.503	0.016	0.073	0.897	0.031	0.025	0.159	0.004	0.882	0.833	0.934	0.028	0	0
0.052	0.098	0.719	0.859	0.074	0.961	0.060	0.056	0.500	0.019	0.072	0.898	0.044	0.025	0.258	0.030	0.709	0.822	0.857	0.025	1	1
0.067	0.166	0.752	0.860	0.074	0.974	0.059	0.058	0.504	0.011	0.066	0.898	0.124	0.023	0.281	0.113	0.861	0.832	0.919	0.027	0	0
0.051	0.072	0.745	0.868	0.076	0.972	0.108	0.058	0.495	0.023	0.032	0.907	0.003	0.024	0.075	0.055	0.841	0.830	0.914	0.028	1	1
0.055	0.053	0.758	0.871	0.079	0.975	0.076	0.058	0.501	0.013	0.044	0.899	0.001	0.027	0.075	0.031	0.851	0.831	0.907	0.027	1	1

TABLE V
CONFUSION MATRICES

Fold 1		Predicted ND	Predicted D	Fold 2		Predicted ND	Predicted D	Fold 3		Predicted ND	Predicted D	Fold 4		Predicted ND	Predicted D	Fold 5		Predicted ND	Predicted D
Actual	ND	192	3	Actual	ND	185	7	Actual	ND	180	6	Actual	ND	172	8	Actual	ND	195	1
	D	15	168		D	9	178		D	7	186		D	7	192		D	11	172

is verified by the testing set. Furthermore, in order to fend off the randomness in selecting testing samples, a fivefold cross validation process is deployed to measure the classifier performance. Using this cross validation process, the whole dataset is separated into five mutually exclusive folds. In each run, one data fold served as testing cases; meanwhile, the other data folds are utilized to establish the classification model. The classifier performance can be evaluated via the average predictive results of the five data folds. Therefore, the bias in data sampling can be avoided and the predictive capability of each prediction model can be measured accurately. The detailed prediction results of the FICDP for the first testing data fold is illustrated in Table IV. In Table IV, Y_A and Y_P denote the actual and predicted output, respectively, an output of 1 indicates a default sample, and an output of 0 represents a nondefault sample. In addition, the confusion matrices of the FICDP's prediction results are shown in Table V.

To better demonstrate the capability of the new approach, its performance is compared to results obtained from the FKNC [15], the SVM [54], the MDA [55], the LR [56], and the naive Bayesian classifier (NBC) [57]. For the FKNN algorithm, the neighboring size *k* is allowed to vary between 1 and 30; additionally, this parameter is selected via a five-fold cross validation process based on the training cases. The fuzzy strength parameter (*m*) in the FKNN algorithm is set to be 2, as recommended by Keller *et al*. [15]. For the SVM algorithm, it is required to specify the regularization parameter (*C*) and the radial basis function kernel parameter (σ). As suggested by [58], the parameters of the SVM is often set as follows: $C = 1$, and $\sigma = 1/D$ where D is the dimension of the input pattern. Furthermore, in this experiment, we also investigate the performance of the SVM when its hyperparameters are tuned by the FA optimization with the same manner as the FICDP. Herein, the FA-tuned SVM prediction model is denoted as the FA-SVM.

TABLE VI
RESULT COMPARISON (AUC)

Prediction Models	Data Folds					Average
	1	2	3	4	5	
FICDP	0.95	0.96	0.97	0.96	0.97	0.96
FA-SVM	0.93	0.93	0.95	0.92	0.96	0.94
SVM	0.91	0.93	0.90	0.91	0.93	0.92
FKNN	0.88	0.94	0.90	0.91	0.92	0.91
LR	0.85	0.85	0.91	0.93	0.87	0.88
MDA	0.87	0.81	0.84	0.83	0.90	0.85
NBC	0.79	0.74	0.75	0.75	0.85	0.78

To benchmark model performances, the AUC is employed. It is worth reminding that an AUC is a portion of the area of the unit squares, its value will always be between 0.0 and 1.0 [45]. It is worth noticing that higher the AUC value better is the model prediction performance. The experimental result of the cross validation process is reported in Table VI. The average AUC values of the FICDP, FA-SVM, SVM, FKNC, LR, MDA, and NBC algorithms are 0.96, 0.94, 0.92, 0.91, 0.88, 0.85, and 0.78, respectively. Moreover, it evidences that the proposed FICDP outperforms other benchmark methods in all data folds. Thus, as can be proved from the experiment, the proposed default prediction is capable of delivering the most desirable outcome in terms of the AUC value.

Based on the experimental results, the instanced-based learning framework is best suited for predicting contractor default prediction in the construction industry. It can be seen that utilizing the information obtained from the nearest patterns of company financial records can help to enhance the forecasting performance. Moreover, the integration of the SMOTE algorithm and this machine learning technique can effectively solve the

class-imbalanced distribution of the default prediction problem. Moreover, since the FICDP has successfully generalized a decision boundary that classifies default and nondefault companies, the proposed model can be employed to forecast performances of construction firms in the future. These facts convincingly demonstrate that the proposed FICDP is a promising alternative for project owners and managers to cope with the contractor default prediction problem.

VI. Conclusion

This study has introduced and verified a novel default prediction model based on various AI techniques. The proposed method, named as FICDP, is established by hybridization of the FKNC, SMOTE, and FA algorithms. The FICDP utilizes the FKNC as a classifier to predict the default/nondefault status when the information of the construction contractor is provided. The SMOTE is integrated into the AI model to solve the imbalance classification problem. Moreover, the FA swarm intelligence is employed to determine the optimal set of model parameters. Hence, the proposed prediction model can be applied easily by practical project managers/owners since it can operate independently without effort in the parameter setting. To verify the performance of the FICDP, benchmarking approaches including the FA-SVM, SVM, FKNC, LR, NBC, and MDA algorithms are used as benchmark approaches. The experiment has shown that the proposed FICDP has achieved the most accurate prediction performance. Thus, the newly proposed method is a promising alternative to help decision makers to deal with the contractor default prediction problem.

References

[1] H. P. Tserng, G.-F. Lin, L. K. Tsai, and P.-C. Chen, "An enforced support vector machine model for construction contractor default prediction," *Autom. Construction*, vol. 20, pp. 1242–1249, 2011.

[2] K. C. Lam, E. Palaneeswaran, and C.-Y. Yu, "A support vector machine model for contractor prequalification," *Autom. Construct.*, vol. 18, pp. 321–329, 2009.

[3] I. M. Horta and A. S. Camanho, "Company failure prediction in the construction industry," *Expert Syst. Appl.*, vol. 40, pp. 6253–6257, Nov. 15, 2013.

[4] R. J. Mason and F. C. Harris, "Predicting company failure in the construction industry," *Proc. Inst. Civil Eng.*, vol. 66, pp. 301–307, 1979.

[5] A. F. Abidali and F. Harris, "A methodology predicting failure in the construction industry," *Construct. Manage. Econ.*, vol. 3, pp. 189–196, 1995.

[6] G. Hall, "Factors distinguishing survivors from failures amongst small firms in the UK construction sector," *J. Manage. Stud.*, vol. 31, pp. 737–760, 1994.

[7] A. Koksal and D. Arditi, "Predicting construction company decline," *J. Construct. Eng. Manage.*, vol. 130, pp. 799–807, 2004.

[8] Y. Huang, "Prediction of contractor default probability using structural models of credit risk: An empirical investigation," *Construct. Manage. Econ.*, vol. 27, pp. 581–596, 2009.

[9] O. S. Al-Sobiei, D. Arditi, and G. Polat, "Predicting the risk of contractor default in Saudi Arabia utilizing artificial neural network (ANN) and genetic algorithm (GA) techniques," *Construct. Manage. Econ.*, vol. 23, pp. 423–430, May 1, 2005.

[10] M.-Y. Cheng, N.-D. Hoang, and Y.-W. Wu, "Hybrid intelligence approach based on LS-SVM and differential evolution for construction cost index estimation: A Taiwan case study," *Autom. Construct.*, vol. 35, pp. 306–313, 2013.

[11] O. Al-Sobiei, D. Arditi, and G. Polat, "Managing owner's risk of contractor default," *J. Construct. Eng. Manage.*, vol. 131, pp. 973–978, 2005.

[12] K. Lam, M. Lam, and D. Wang, "Efficacy of using support vector machine in a contractor prequalification decision model," *J. Comput. Civil Eng.*, vol. 24, pp. 273–280, 2010.

[13] K. C. Lam and C. Y. Yu, "A multiple kernel learning-based decision support model for contractor pre-qualification," *Autom. Construct.*, vol. 20, pp. 531–536, 2011.

[14] M.-Y. Cheng and N.-D. Hoang, "Interval estimation of construction cost at completion using least squares support vector machine," *J. Civil Eng. Manage.*, vol. 20, pp. 223–236, 2013.

[15] J. M. Keller, M. R. Gray, and J. A. Given, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.

[16] S.-T. Li and H.-F. Ho, "Predicting financial activity with evolutionary fuzzy case-based reasoning," *Expert Syst. Appl.*, vol. 36, pp. 411–422, Jan.,2009.

[17] M. Cheng and N. Hoang, "Groutability estimation of grouting processes with microfine cements using an evolutionary instance-based learning approach," *J. Comput. Civil Eng.*, vol. 28, p. 04014014, 2014.

[18] H.-L. Chen, C.-C. Huang, X.-G. Yu, X. Xu, X. Sun, G. Wang, and S.-J. Wang, "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach," *Expert Syst. Appl.*, vol. 40, pp. 263–271, Jan. 2013.

[19] H.-L. Chen, B. Yang, G. Wang, J. Liu, X. Xu, S.-J. Wang, and D.-Y. Liu, "A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method," *Knowl.-Based Syst.*, vol. 24, pp. 1348–1359, Dec. 2011.

[20] S. T. Ng, M. Skitmore, and K. F. Wong, "Using genetic algorithms and linear regression analysis for private housing demand forecast," *Building Environ.*, vol. 43, pp. 1171–1184, 2008.

[21] C.-H. Wu, G.-H. Tzeng, and R.-H. Lin, "A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression," *Expert Syst. Appl.*, vol. 36, pp. 4725–4735, 2009.

[22] M.-Y. Cheng, H.-S. Peng, Y.-W. Wu, and T.-L. Chen, "Estimate at completion for construction projects using evolutionary support vector machine inference model," *Autom. Construct.*, vol. 19, pp. 619–629, 2010.

[23] S. Kiranyaz, T. Ince, A. Yildirim, and M. Gabbouj, "Evolutionary artificial neural networks by multi-dimensional particle swarm optimization," *Neural Netw.*, vol. 22, pp. 1448–1462, 2009.

[24] S. S. Gilan, H. B. Jovein, and A. A. Ramezanianpour, "Hybrid support vector regression—Particle swarm optimization for prediction of compressive strength and RCPT of concretes containing etakaolin," *Construct. Building Mater.* vol. 34, pp. 321–329, 2012.

[25] M.-Y. Cheng and N.-D. Hoang, "Groutability prediction of microfine cement based soil improvement using evolutionary LS-SVM inference model," *J. Civil Eng. Manage.*, vol. 20, pp. 1–10, 2014.

[26] F.-K. Wang and T. Du, "Implementing support vector regression with differential evolution to forecast motherboard shipments," *Expert Syst. Appl.*, vol. 41, pp. 3850–3855, Jun. 15, 2014.

[27] X.-S. Yang, *Firefly Algorithm*. Bristol, U.K.: Luniver Press, 2008.

[28] T. Xiong, Y. Bao, and Z. Hu, "Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting," *Knowl.-Based Syst.*, vol. 55, pp. 87–100, Jan. 2014.

[29] B. Amiri, L. Hossain, J. W. Crawford, and R. T. Wigand, "Community detection in complex networks: multi–objective enhanced firefly algorithm," *Knowl.-Based Syst.*, vol. 46, pp. 1–11, Jul. 2013.

[30] P. Mandal, A. U. Haque, M. Julian, A. K. Srivastava, and R. Martinez, "A novel hybrid approach using wavelet, firefly algorithm, and fuzzy ARTMAP for day-ahead electricity price forecasting," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1041–1051, May 2013.

[31] A. K. Fard and T. Niknam, "Optimal stochastic capacitor placement problem from the reliability and cost views using firefly algorithm," *IET Sci., Meas. Technol.*, vol. 8, pp. 260–269, 2014.

[32] R. Batuwita and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets," in *Proc. Int. Joint Conf. Neural Netw.*, 2010, pp. 1–8.

[33] R. Barandela, J. S. Sánchez, V. Garcia, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognit.*, vol. 36, pp. 849–851, 2003.

[34] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inform. Sci.*, vol. 250, pp. 113–141, Nov. 20, 2013.

[35] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," *Appl. Soft Comput.*, vol. 20, pp. 15–24, Jul. 2014.

[36] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, pp. 3456–3466, 2011.
[37] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inform. Sci.*, vol. 291, pp. 184–203, Jan. 10, 2015.
[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[39] I. Fister, I. Fister, Jr, X.-S. Yang, and J. Brest, "A comprehensive review of firefly algorithms," *Swarm Evol. Comput.*, vol. 13, pp. 34–46, Dec. 2013.
[40] X.-S. Yang, *Nature-Inspired Optimization Algorithms*. New York, NY, USA: Elsevier, 2014.
[41] A. Baykasoğlu and F. B. Ozsoydan, "An improved firefly algorithm for solving dynamic multidimensional knapsack problems," *Expert Syst. Appl.*, vol. 41, pp. 3712–3725, 2014.
[42] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decision Support Syst.*, vol. 48, pp. 191–201, Dec. 2009.
[43] Z.-Q. Zhao, "A novel modular neural network for imbalanced classification problems," *Pattern Recognit. Lett.*, vol. 30, pp. 783–788, 1 Jul., 2009.
[44] G. Weiss and Y. Tian, "Maximizing classifier utility when there are data acquisition and modeling costs," *Data Mining Knowl. Discovery*, vol. 17, pp. 253–282, Oct. 1, 2008.
[45] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, jun.//2006.
[46] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, pp. 1145–1159, Jul. 1997.
[47] H. Jin and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.
[48] I. D. Dichev, "Is the risk of bankruptcy a systematic risk?," *J. Finance*, vol. 53, pp. 1131–1147, 1998.
[49] P. Brockman and H. J. Turtle, "A barrier option framework for corporate security valuation," *J. Financial Econ.*, vol. 67, pp. 511–529, Mar. 2003.
[50] A. F. Abidali and F. Harris, "A methodology for predicting company failure in the construction industry," *Construct. Manage. Econ.*, vol. 13, pp. 189–196, May 1, 1995.
[51] G. Severson, J. Russell, and E. Jaselskis, "Predicting contract surety bond claims using contractor financial data," *J. Construct. Eng. Manage.*, vol. 120, pp. 405–420, Jun. 1, 1994.
[52] J. Russell and H. Zhai, "Predicting contractor failure using stochastic dynamics of economic and financial variables," *J. Construct. Eng. Manage.*, vol. 122, pp. 183–191, 1996.
[53] R. Kangari, F. Farid, and H. M. Elgharib, "Financial performance analysis for construction industry," *J. Construct. Eng. Manage.*, vol. 118, pp. 349–361, 1992.
[54] V. N. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.
[55] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA:, Springer, 2006.
[56] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. San Diego, CA, USA: Academic, 2009.
[57] R. O. Duda, P. E. Hart, and D. G. Stock, *Pattern Classification*, 2nd Ed. Hoboken, NJ, USA: Wiley, 2001.
[58] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2010.

Min-Yuan Cheng received the Ph.D. degree in civil engineering and construction management from the University of Texas at Austin, Austin, TX, USA, in 1999.

He is currently a Distinguished Professor of Construction Management, Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taiwan. His research interests include construction process re-engineering, Geographic Information System (GIS), automation and E-business in construction, and artificial intelligence applications in construction management.

Dr. Cheng is an Editorial Board Member of the *Journal of Automation in Construction* and the *Journal of Civil Engineering and Management*.

Nhat-Duc Hoang received the Ph.D. degree in construction management from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2013.

He is currently a Researcher and Lecturer at the Institute of Research and Development and Faculty of Civil Engineering, Duy Tan University, Vietnam. His research interests include applications of artificial intelligence in construction engineering and management.