Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis

Article *in* IEEE Transactions on Engineering Management · February 2020

DOI: 10.1109/TEM.2020.2972078

CITATIONS
72

READS
1,925

1 author:

Haytham Elmousalami
Zagazig University
27 PUBLICATIONS **454** CITATIONS

SEE PROFILE

Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis

Haytham H. Elmousalami

***Abstract*—Developing a reliable parametric cost model at the conceptual stage of the project is crucial for project managers and decision makers. Existing methods, such as probabilistic and statistical algorithms have been developed for project cost prediction. However, these methods are unable to produce accurate results for conceptual cost prediction due to small and unstable data samples. Artificial intelligence (AI) and machine learning (ML) algorithms include numerous models and algorithms for supervised regression applications. Therefore, a comparative analysis for AI models is required to guide practitioners to the appropriate model. The article focuses on investigating 20 AI techniques which are conducted for conceptual cost modeling, such as fuzzy logic model, artificial neural networks, multiple regression analysis, case-based reasoning, hybrid models, such as genetic fuzzy model, and ensemble methods such as scalable boosting trees (XGBoost) and random forest. Field canals improvement projects (FCIPs) are used as an actual case study to analyze the performance of the applied ML models. Out of 20 AI techniques, the results show that the most accurate and suitable method is XGBoost with 9.091% and 0.929 based on mean absolute percentage error and adjusted R^2, respectively. Nonlinear adaptability, handling missing values and outliers, model interpretation, and uncertainty have been discussed for the 20 developed AI models. In addition, this study presents a publicly open dataset for FCIPs to be used for future models' validation and analysis.**

***Index Terms*—Artificial intelligence (AI), conceptual cost, ensemble methods, evolutionary fuzzy rules generation, machine learning (ML), parametric cost model, and XGBoost.**

I. Introduction

CONCEPTUAL cost estimate occurs at 0–2% of the project completion where limited information about the project is available with a high level of uncertainty and unknown risks [39]. Conceptual cost prediction is considered one of the core criteria in the project's decision making at the early stages of the project. Cost model estimates the conceptual cost of the project. Conceptual cost value is one of the key criteria to take a financial decision about the project. Moreover, conceptual cost value is used for risk analysis and feasibility study of the project [29].

The estimating must be completed within a limited time period. Therefore, the precise conceptual cost estimate is a challengeable task for cost engineers, project managers, and decision makers [43]. Capacity factored model, analog model (near neighbor), and parametric model are conducted to perform such conceptual estimate where its accuracy varies from −50% to +100% [1]. Parametric cost estimate performs more accurate results than capacity factored model and analog model where parametric cost estimate deeply construct cost estimating relationships (CERs) between cost and cost predictors [1], [69]. Parametric cost modeling is developing a model based on key cost drivers extracted from experts' experience or the collected past cases by conducting statistical analyses [58].

The main motivations to automate cost estimation are as follows.

1) Quantity survey is a time and effort consuming process [74].
2) Cost estimation may prone to human errors during estimation or personal judgment where biases and inaccuracy can exist.
3) Highly accurate and reliable tool is required for project managers and decision makers to evaluate the conceptual cost of the proposed project.

Artificial intelligence (AI) involves powerful algorithms to automate cost estimate with high precision based on collected project data. However, the accuracy of cost prediction is a major criterion in the success of any construction project, where cost overruns are a critical unknown risk, especially with the current emphasis on tight budgets. Moreover, cost overruns can lead to the cancellation of the project [1], [34]. Therefore, improving prediction accuracy is the main requirement in developing the cost model. Small data size, missing data values, maintaining uncertainty, computational complexity, and model interpretation are the key challenges during prediction modeling.

II. Literature Review

Serval previous literature have discussed applications of AI and machine learning (ML) models for construction cost estimate. Marzouk and Elkadi [61] have applied ANNs where the mean absolute percentage error (MAPE) for test sets was 21.18%. Williams and Gong [82] have built a stacking ensemble

Manuscript received April 26, 2019; revised September 7, 2019 and October 25, 2019; accepted January 21, 2020. Review of this manuscript as arranged by Department Editor P. Hung.

The Author is with the Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt, and also with the General Petroleum Company, Cairo 11311, Egypt (e-mail: haythamelmousalami2014@gmail.com).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEM.2020.2972078

learning and text mining to estimate the cost overrun using the project contract document where the accuracy was 44%. However, Peurifoy and Oberlender [68] have defined 20% as an acceptable limit for the conceptual cost estimate based on MAPE. Therefore, the key gap of these studies Marzouk and Elkadi [61]; Williams and Gong [82] is prediction accuracy.

A semilog regression model has performed to develop cost models for residential building projects in German with a prediction accuracy of 7.55% [79]. Based on 92 building projects, artificial neural networks (ANNs) and supportive vector machine (SVM) have been used to predict cost and schedule success at the conceptual stage. Such a model has a prediction accuracy of 92% and 80% for cost success and schedule success, respectively [81]. Based on 657 building projects in Germany, a multistep ahead approach is conducted to increase the accuracy of the model's prediction [28]. Fan *et al*. [32] have developed a decision tree (DT) approach for investigating the relationship between house prices and housing characteristics. Monte Carlo simulation and a multiple linear regression model have been developed as a benchmark model to evaluate the model's performance where the MAPE was 7.56. Wang and Ashuri [92] have developed a highly accurate model based on random tree ensembles to predict construction cost index where the model's error has reached 0.8%. Chou and Lin [23] have established an ensemble learning model of ANNs, SVM, and DT for predicting the potential for disputes in public-private partnership with the accuracy of 84%. Analytic hierarchy process has incorporated into case-based reasoning (CBR) to build a reliable cost estimation model for highway projects in South Korea [48]. The previous studies have applied ANNs, CBR, and SVM.

ANN model has been conducted to the early cost estimate of building projects for reinforced concrete buildings with acceptable performance [6]. CBR has been proposed for estimation of the preliminary costs of sports field construction based on 16 predictors using 143 construction projects. Different calculations were conducted to formulate the case similarity based on quantitative and qualitative data where the final total error was 14% at the early stage [54]. Prediction performance of a cost prediction model has been improved by 17.23% and 4.39% for business facilities model and multifamily housings model, respectively.

Based on more than 1,400 projects, a multilayer of ensemble methods has been developed for forecasting the unit price bids of resurfacing highway projects [18]. Wang and Ashuri [92] have applied a random tree model for construction cost index prediction. Williams and Gong [82] have built a stacking ensemble learning and text mining to estimate the cost overrun using the project contract document where the accuracy was 44%. Building information modeling (BIM) can automate cost estimation process and improve inaccuracies where new rules of measurement for cost estimation can be extracted for automatic cost estimate based on a 4-D BIM modeling software [49].

Arabzadeh *et al*. [8] have developed ANNs, regression and hybrid models for cost estimation of spherical storage tanks. The results indicated that ANN was more accurate than a hybrid regression model and hybrid ANNs was more accurate than single ANNs. Linear and multiple regression models have been counted to predict the preliminary estimate of road projects in Nigeria at the early stage [63]. However, the whole collected data set was only 50 for seven predictors where it is not sufficient data size to train regression models. Zhang *et al*. [91] have converted the time series model into a graph to forecast the construction cost index where the application showed its ability to provide more accurate estimations.

However, these algorithms have several limitations. ANNs have limited accuracy with small training data set. CBR, ANNs, and SVM cannot handle missing data values or noisy data. In addition, these techniques are black box nature where no results interpretations are possible. Moreover, the main gap of these studies is developing deterministic predictive models without taking uncertainty nature into account where adding uncertainty nature to the predicted values improve the quality and reliability of the developed models [87], [88].

A. Research Gaps and Objectives

Automated cost models are prone to many ML problems, such as overfitting issues and hyper-parameter selection. It is necessary to develop more than one cost prediction model, such as regression, ANNs, FL, or CBR. As a result, the researcher aims to compare the results of the developed ML models and set evaluation criteria to select the optimal model. In addition, the comparisons of the developed models enhance the quality of cost estimate and the decisions based on it [5].

The conceptual cost estimate is conducted under uncertainty. Therefore, this study recommends using fuzzy theory, such as fuzzy logic (FL) and to develop a hybrid model based on FL such as genetic fuzzy model to obtain uncertainty for the developed model and produce a more reliable performance [29]. However, these uncertainty-based models such as fuzzy and genetic fuzzy models must compared to other ML algorithms to select the most accurate model.

Therefore, this study aims to present a comprehensive comparison analysis for AI modeling to overcome the previous limitations. Predictive modeling consists of two main stages: Feature engineering and model development. Feature engineering is to select the key cost drivers as input to compute the project cost. Model development is to apply statistical learning techniques to know the pattern between the key cost drivers and project cost.

B. Feature Engineering

Parametric cost modeling is to develop a model based on logical or statistical relations of the key cost drivers extracted by conducting qualitative techniques [30] or statistical analyses, such as factor analysis [61] or stepwise regression technique [31]. Marzouk and Elkadi [61] identified cost drivers that influence construction costs of water treatment plants. Cost drivers have been determined through descriptive statistics ranking and EFA. Principal component analysis with varimax rotation through five iterations were used to minimize multicollinearity problem. A total of 33 variables were reduced to eight components while using Cattell's Scree test reduced variables to four components. Woldesenbet and Jeong [84] have conducted the factor analysis algorithms of a covariance and correlation matrix

to investigate the significance and correlation of critical factors affecting the preliminary cost of roadway projects. Alroomi *et al*. [4] identified 23 core estimating competencies classified into skills, knowledge, and personal attributes and also quantified the degree where new estimators each lack competency. The factor analysis has grouped these 23 competencies into seven different factors by using the factor analysis method.

Stoy [80] developed conceptual cost models for German residential building project. Historical data were randomly sampled from the building cost information centre. A total of 75 residential projects have been sampled. Multicollinearity and singularity problems have been detected and eliminated where the most significant predictors were compactness, the percentage of openings, and height of the building for the cost of external walls. These parameters were determined by a backward regression method. Lowe and Emsley [57] described the development of linear regression models to predict the construction cost of buildings, based on 286 sets of data collected in the United Kingdom. Both forward and backward stepwise regression analyses were performed to produce a total of six models. A total of 41 independent variables have been identified and classified either as project strategic, site-related or design related.

Yang [85] presented a general method to incorporate correlations between cost elements in the process of cost estimation. Yang [85] proposed a simulation-based method to estimate project cost while considering correlations between cost elements where it can automatically adjust an infeasible correlation matrix into a close and feasible one very efficiently. The proposed method has first checked the feasibility of the correlation matrix, adjusts it if necessary, then has used the correlations to generate correlated multivariate random vectors to generate outcomes of the cost elements. The method was applied to a full data set of 216 British office buildings. The application result indicated that the impact of correlations was significant and may cause serious problems if neglected. Ranasinghe [93] used the correlation matrix for selecting the input project cost variables based on 70 German residential properties. Stoy *et al*. [94] used a series of independent variables for early estimation of building construction cost of residential buildings by regression analysis. These variables serve as cost drivers of a project. As illustrated by the literature survey, most studies conducted factor analysis, regression analysis, or correlation matrix to select the key cost drivers.

C. Model Development

Once key cost drivers have been identified, the model development stage can be started. Many previous studies have applied AI techniques and ML models. BIM can feed data for cost estimation where a predictive ML model such as the regression model or ANNs can predict the project's cost on a macro level [44]. ANN has been applied for cost estimation of sports fields where the general applicability of ANNs model has been investigated [45].

Fuzzy theory can be conducted to handle uncertainty concept to prediction modeling [87], [88]. Based on 568 Towers, a four-input fuzzy clustering model and sensitivity analysis are conducted for estimating telecommunication towers with acceptable MAPE [60]. Shreenaath *et al*. [77] have conducted a statistical fuzzy approach for prediction of construction cost overrun. The FL model is developed for satellite cost estimation. Such model works as a fuzzy expert tool for cost prediction based on two input parameters [46]. However, these studies have developed fuzzy systems without mentioning the method of fuzzy rules generation or the fuzzy rules has been developed based on experts' experience. Determining the fuzzy rules is the main gap of the previous studies. Therefore, a new trend evolves to solve this problem such as developing hybrid fuzzy modeling for the cost estimate purposes such as evolutionary-fuzzy modeling.

Zhu *et al.* [95] have conducted an evolutionary fuzzy neural network model for cost estimation based on 18 examples and two examples for training and testing, respectively. The previous study has an insufficient sample size for model training where Green [36] has recommended that $50 + 8k$ may be the minimum sample size, where k is the number of predictors. GA is conducted for model optimization and to avoid sinking into local minimum results. Cheng and Roy, [20] have developed a hybrid AI system based on SVM, FL, and GA for decision making construction management. The system has applied the FL to handle uncertainty to the system, SVM to map fuzzy inputs and outputs, and GA to optimize the FL and SVM parameters.

Zhai *et al.* [90] have created an improved fuzzy system which is established based on fuzzy c-means to solve the problem of fuzzy rules generation. This model has produced better results for scientific cost prediction. Cheng *et al*. [21] have incorporated computation intelligence models, such as ANNs, FL, and EA to make a hybrid model which improves the prediction accuracy. FL is used for fuzzification and defuzzification for inputs and outputs, respectively. GA is utilized for optimizing the parameter of the model such as NN layer connections and FL membership. As a result, an evolutionary fuzzy neural model has been developed for conceptual cost estimation for building projects.

III. Research Methodology

Based on the literature survey, there is no comprehensive comparison for different AI models for conceptual cost modeling. The main objective of this article is to evaluate the prediction accuracy of different AI models to produce the most accurate cost prediction model. Moreover, this article aims to present a comprehensive performance comparison for AI model to guide researchers and practitioners during conceptual cost modeling. The scope of this study focuses on the most common AI techniques such as SVM, FL model, ANNs, multiple regression analysis (MRA), CBR, hybrid models, and ensemble learning methods, DT, random forest (RF), Adaboost, scalable extreme gradient boosting machines (XGBoost), and hybrid models such as genetic fuzzy model. This article consists of five steps as follows as illustrated in Fig. 1.

1) The first step is to review the past literature to know the past practices for conceptual cost modeling.
2) The second step includes data collection of real historical cases of Field canals improvement projects (FCIPs). The

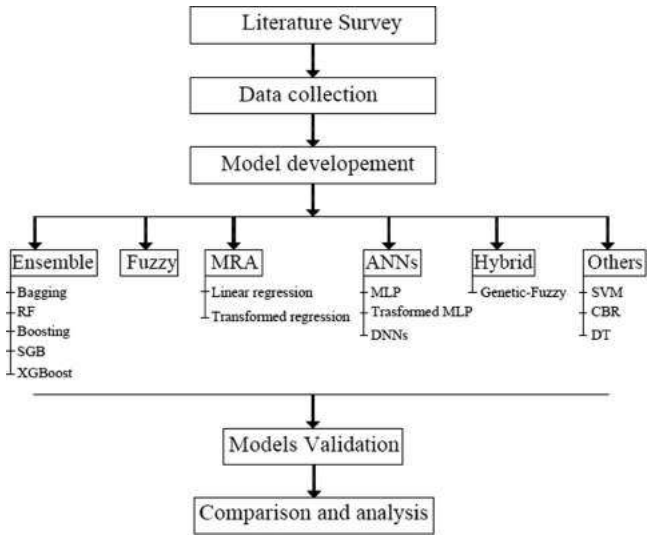

Fig. 1. Research methodology.

data have been quantitatively collected based on the contract information for each project where "Data collection and feature selection" section will discuss in details.

3) The third step includes a model development based on AI models where a total of 20 predictive models have been built.
4) The fourth step is models' validation to select the most accurate model using the evaluation techniques such as coefficient of determination (R2) and MAPE.
5) The fifth step is to analyze the results and conduct a sensitivity analysis to determine the contribution of each key parameter on the total cost of FCIPs.

IV. Application to FCIPs

AI can automatically develop the relations among cost drivers and the project's cost where the final prediction error can be minimized. Therefore, AI can diminish human interventions to estimate the project cost. Moreover, the automated parametric model needs a few cost drivers as inputs to predict the final cost without quantity survey using low computational time and memory [31]. In this section, the selected AI techniques are applied to the conceptual cost prediction of FCIPs in Egypt as an actual case study.

A. Case Background

FCIPs are one of the main projects in irrigation improvement projects in Egypt. The strategic aim of these projects is to save fresh water, facilitate water usage and distribution among stakeholders and farmers. To finance this project, conceptual cost models are important to accurately predict preliminary costs at the early stages of the project [31], [72].

B. Data Collection and Feature Selection

Based on contracts information, cost drivers of FCIPs can be represented through a total of 17 parameters [31] as shown in Table I. Data preprocessing includes data normalization, cleaning, and transformation. Once the inputs (key cost drivers) and the output (conceptual construction cost of FCIP) were identified, relevant data were collected to build the parametric cost model. The quantity and quality of the collected instances are significant for the conceptual estimate that affects the accuracy of the developed model [13].

TABLE I
Collected Parameters of FCIPs [31]

Notation	Variables
P_1	FCIP area served
P_2	Average area of area served sections
P_3	Pipeline total length
P_4	Equivalent diameter of the pipeline
P_5	Duration of the FCIP
P_6	Irrigation valves number
P_7	Pressure relief valves number
P_8	Sump size
P_9	Pump house size
P_{10}	max discharge capacity
P_{11}	Electrical pump discharge
P_{12}	Diesel pump discharge
P_{13}	Orientation of the improved canal
P_{14}	Year of construction
P_{15}	Rice existence
P_{16}	Intake existence
P_{17}	Parallel canal

These 17 parameters can be filtered to be entered to ML models. Therefore, Elmousalami *et al*. [30] have conducted qualitative approaches such as Fuzzy Delphi method and fuzzy analytical hierarchy process to rank the cost drivers. Moreover, Elmousalami *et al*. [31] have developed a quantitative hybrid approach based on both Pearson correlation and stepwise regression to filter the key cost drivers. Accordingly, the final key cost drivers were area served (P_1), pipeline total length (P_2), the number of irrigation valves (P_3), and construction year (P_4). Accordingly, a total of 144 FCIPs during 2010 and 2015 have been collected.

For validation purposes, this collected sample has randomly branched into a training sample (111 instances) and a testing sample (33 instances). The training sample in the present case study is 111 instances which would be sufficiently acceptable to train reliable ML models where Green [36] concluded that [50 + 8∗N] is the minimum sample size, and N is the number of independent variables (key cost drivers).

V. AI Techniques Developments

AI techniques are aspects of human knowledge and computational adaptively to become more vital in system modeling than classical mathematical modeling [11]. Based on AI, an intelligent system can be developed to produce consequent outputs and actions depending on the observed inputs and outputs of the system [12], [78]. AI and ML are general purpose techniques which can be applied for a wide range of applications [83]. For example, ANNs or DT can be used for cost prediction in

the construction industry or for DNA sequence prediction in bioinformatics [27]. Similarly, the AI models in this study can be applied in the abroad area of construction projects where modeling methodology can valid for different projects types. The AI models can be categorized into single AL models and ensemble AI models

A. Single AL Models

1) Case-Based Reasoning: CBR is a sustained learning and incremental approach that solves problems by searching the most similar past case and reusing it for the new problem situation [2]. Therefore, CBR mimics a human problem solving [51], [73]. CBR is a cyclic process of learning from past cases to solve a new case. The main processes of CBR are retrieving, reusing, revising, and retaining. The retrieving process is solving a new case by retrieving the past cases. The case can be defined by key attributes. Such attributes are used to retrieve the most similar case. The reusing process is utilizing the new case information to solve the problem. The revising process is evaluating the suggested solution to the problem. Finally, the retaining process is to update the stored past cases with such a new case by incorporating the new case to the existing case-base [2]. A CBR model is developed to predict the conceptual cost of FCIP based on similarity attribute of the entered case comparable with the stored cases. Once attributes are entered, attributes similarities (AS) can be computed based on the following [47]:

$$\mathrm{AS} = \frac{\mathrm{Min}\,(AV_N, AV_R)}{\mathrm{Max}\,(AV_N, AV_R)} \tag{1}$$

where AS = Attribute Similarity, $\mathrm{AV_N}$ = Attribute value of the newly entered case, $\mathrm{AV_R}$ = Attribute value of the retrieved case. Depending on AS and attribute weight (AW), case similarity (CS) can be computed by (2) [66]. AW is selected by an expert to emphasize the existence and importance of the case attributes.

$$\mathrm{CS} = \frac{\sum_{i=1}^{n} (AS_i * \ AW_i\ )}{\sum_{i=1}^{n} (\ AW_i\ )} \tag{2}$$

where CS is case similarity, AS is the attribute similarity, AW is the attribute weight, and *i* is the number of the attributes (key cost drivers).

2) FL Model: FL is to model human reasoning taking uncertainties possibilities into account where incompleteness, randomness, and ignorance of data are represented in the model [87], [88]. If–Then rules are logical inference statements which are utilized to formulate the FL rules base system as shown example in Fig. 2. Fig. 2 illustrates the fuzzy rules firing process. Each deterministic value converts into fuzzy values using fuzzy membership functions (FMF) to form antecedent rules. Antecedent rules produce consequent rules based on fuzzy rules system and consequent rules converted into deterministic values by fuzzifications method [29].

There are two parameters X_1 and X_2 where $\mu\, X_1 = \{a_1, b_1, c_1, d_1\}$, $\mu\, X_2 = \{a_2, b_2, c_2, d_2\}$, $\mu\, Y = \{a_y, b_y, c_y, d_y\}$ and the fuzzy system consists of two rules as follows:

Rule 1: IF $\mathrm{x_1}$ is $\mathrm{a_1}$ AND $\mathrm{x_2}$ is $\mathrm{c_2}$ THEN y is $\mathrm{a_y}$.

Rule 2: IF $\mathrm{x_1}$ is $\mathrm{b_1}$ AND $\mathrm{x_2}$ is $\mathrm{d_2}$ THEN y is $\mathrm{b_y}$.

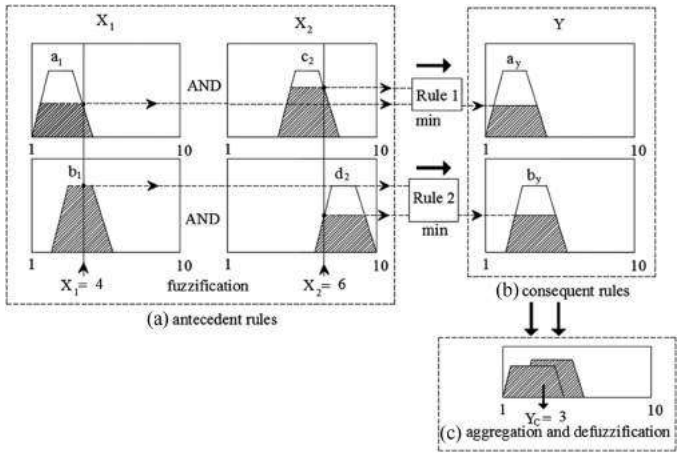

Fig. 2. Fuzzy rules firing: (a) antecedent rules; (b) consequent rules; (c) aggregations; and fuzzifications.

where two inputs are used $\{X_1 = 4, X_2 = 6\}$. Such two inputs intersect with the antecedents MF of the two rules where two consequents rules are produced $\{R_1$ and $R_2\}$ based on minimum intersections.

FMF are triangular, z-shape, trapezoidal, s-shape, sigmoid, and Gaussian. The choice of membership function depends entirely on the problem size and problem type. Triangular fuzzy membership is appropriate to the sample size of the FCIP case study and triangular fuzzy membership need low computational cost [50], [2].

The consequent rules are aggregated based on maximum intersections where the final crisp value is three. The aggregated output for R_i rules is given by

Rule 1: μ R1 = min [μ a1 (x1) and μ c2 (x2)]

Rule 2: μ R2 = min [μ b1 (x1) and μ d2 (x2)]

Y: Fuzzification [max [$\mathrm{R_1}$, $\mathrm{R_2}$]

The first step in the FL model is fuzzification of the four key cost drivers and identifying their MFs. The most critical stage is to develop fuzzy rules base. Experts are consulted to give their experience to develop such rules. As shown in Fig. 3, seven triangle MFs have been used to fuzzify the variables of FCIPs for example, the input variable construction year consists of seven triangle MFs {$\mathrm{MF_1}$, $\mathrm{MF_2}$, $\mathrm{MF_3}$, $\mathrm{MF_4}$, $\mathrm{MF_5}$, $\mathrm{MF_6}$, $\mathrm{MF_7}$}. Accordingly, the number of possible rules equals 7^4 rules. Therefore, there is a need to automatically generate such rules. For the FL model, a total of 190 IF–Then inference rules have been formulated based on the fuzzy designer logic.

3) Genetic-Fuzzy Model: Many approaches exist for evolutionary fuzzy hybridization [7], [65]. Traditionally, an expert is consulted to define such fuzzy rules or the fuzzy designer can use the trial-and-error approach to map the fuzzy rules and MFs. However, such an approach is time-consuming and does not guarantee the optimal set of fuzzy rules. Moreover, the number of fuzzy IF–Then inference rules increase exponentially by increasing the number of inputs, linguistic variables, or a number of outputs. In addition, the experts cannot easily define all required fuzzy rules and the associated MFs. In many

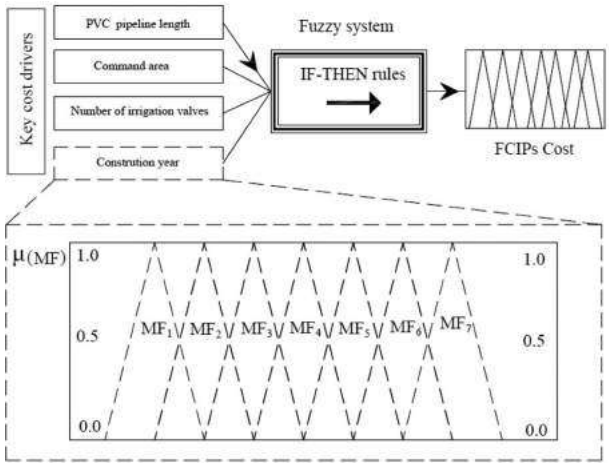

Fig. 3. FL model for FCIPs [29].

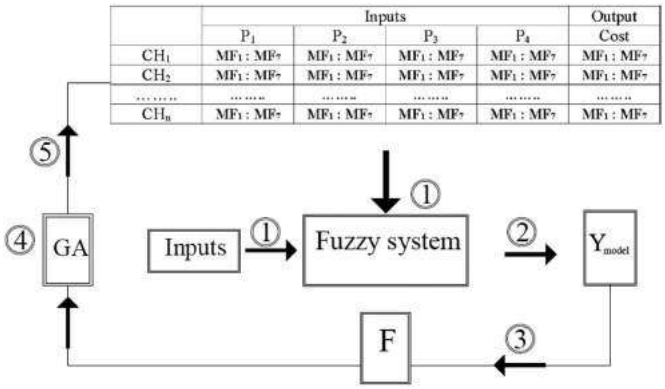

Fig. 4. Process of genetic fuzzy system.

engineering problems, the evolutionary algorithm (EA) has been conducted to automatically develop fuzzy rules and MFs to improve system performance [22], [56].

Genetic-fuzzy model has been developed to optimally generate fuzzy inference rules. The study has applied the genetic algorithm (GA) to optimally select the fuzzy rules where 2401 rules represent the whole possible search space for GA. The formulation of the GA model depends mainly on defining two core terms: A chromosome representation and an objective function. First, based on Michigan approach, the chromosomes represent the fuzzy rules where the number of chromosomes (CH_n) are the number of fuzzy rules. Each chromosome is consisting of five genes where four genes are for the key cost drivers the fifth gene is for the output (the cost of FCIP). Each gene consists of one of the seven membership functions (MF_i) where (*i*) is ranging from one to seven (MF_1:MF_7) as shown in Fig. 4. For example: **IF** {Area served (P_1) is MF_5 **AND** Total length (P_2) is MF_2 **AND** Irrigation valves (P_3) is MF_2 **AND** construction year (P_4) is MF_6} **THEN** {The Cost LE / Mesqa is MF_3}.

Second, the fitness function is problem-dependent where the objective is to enhance the accuracy and quality of the system performance [38]. The fitness function is formulated to minimize the MAPE as (3).

$$F = MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\hat{y}_i}x100 \tag{3}$$

where: (F) is a fitness function, (n) is the number of cases, (i) is the number of the case and $\hat{y}_i$ is the outcome of the model and y_i is the actual outcome. As shown in Fig. 4, the process of the developed model consists of five main steps as follows.

1) An initial population of chromosomes has been identified to represent the initial state of the fuzzy rules. The four key cost drivers have been fed to the fuzzy system.
2) The fuzzy system produces the final output of the system $\hat{y}_i$.
3) The predicted cost $(\hat{y}_i)$ has been fed to fitness function (F) to evaluate the model performance.
4) GA uses the fitness function (F) to evaluate the search process where crossover probability and mutation probability have been set at 0.7 and 0.01, respectively.
5) The new population of fuzzy rules has been produced based on the crossover and the mutation processes to converge the optimal fuzzy rules.

Based on genetic fuzzy model, the study conducted an experimental analysis to select the optimal FMF. Seven FMFs have been applied separately to the genetic fuzzy model based on accuracy (MAPE and R∗2) and computational cost (computational time and memory) as show in Table II. the optimal FMF was Gaussian function where it had the higher accuracy of 12.9 and 0.89 based on MAPE and adjusted R2, respectively. On the other hand, the computational costs were approximately similar where the computational memory was ranging from 9.06 to 13.57 Mb and the computational memory was ranging from 3.6 to 8.71 s. Triangular function was the optimal function based on the computational complexity. However, the decision-maker needs the model with the higher accuracy and acceptable computational cost. Therefore, the final optimal FMF was Gaussian function.

4) Support Vector Machines: SVM is a nonparametric supervised ML algorithm that can be applied for regression and classification problems [8 0]. The study has applied the radial base function (RBF) as a kernel for the supportive vector regression model. A positive slack variable (ξ) is added to handle the nonlinearity of the data in the following [24]:

$$y_i\,(W.X_i + b) \geq 0 - \ \xi.i = 1, 2, 3, \ldots\ldots\ m \tag{4}$$

The objective is to minimize misclassifications cases by optimizing the margin and hyperplanes distance as follows:

$$Min \quad \sum_{i=0}^{i=m}\frac{1}{2}w.w^T + C\sum_{i=0}^{i=m}\xi_i. \tag{5}$$

For *i = 1, 2, 3, …..*, *m* where *m* is the number of cases.

5) Decision Trees: DT is a supervised ML model that divides the cost data into hierarchical rules on each tree node by a repetitive splitting algorithm [10], [17]. Classification and regression trees (CART) model is a DT model that can be applied for both regression (continuous variables) and classification (categorical variables) applications [71]. CART has been developed to the FCIPs data. The features of the tree are the key cost drivers and

TABLE II
COMPUTATION OF DIFFERENT FMF

FMF	MAPE %	R2	R*2	Computational time (Seconds)	Computational memory (Megabits)
Triangular	14.700	0.863	0.857	3.6	9.06
z-shape	16.6	0.842	0.83	6.4	10.1
Trapezoidal	15.800	0.81	0.807	5.3	10.3
S-shape	15.300	0.79	0.79	8.71	11.78
Sigmoid	14.95	0.863	0.86	5.3	10.3
Gaussian	**12.900**	**0.893**	**0.890**	**7.7**	**13.57**
Generalized bell	13.200	0.871	0.881	6.3	12.24

the terminal tree nodes (leaf nodes) are continues values of the project cost.

6) MRA and Transformed Regressions: Elmousalami *et al*. [31] have developed five regression models: Standard linear regression, quadratic model, reciprocal model, semilog model, and power model. The most accurate model is the quadratic model where the quadratic model is a dependent variable transformation by taking the square root (Sqrt). The regression model consists of four key cost drivers as independent variables and (Y) represent FCIP cost per field canal as the dependent predictor. The quadratic regression model is formulated as follows:

$$(Y)^{0.5} = -37032.81 + 2.21^{*}P_1 + 0.1691^{*}P_2 + 2.265^{*}P_3 + 18.594^{*}P_4. \quad (6)$$

7) ANNs and DNNs: ANNs is a computational method that is inspired by neuron cells. The major advantage of ANNs is their ability to fit nonlinear data [78]. Elmousalami *et al*. [31] have developed three ANNs models with structure (4-5-0-1) where four represents the number of inputs (four key cost drivers), five represents the number of hidden nodes in the first hidden layer, zero means no second hidden layer used, and one represents one node to produce the total cost of the FCIPs. The first model is the untransformed model whereas the second model is transformed by the square root of the project cost. The third model is transformed by the natural log of the project cost. The type of training is batch, the learning algorithm is the scaled conjugate gradient and the activation function is hyperbolic tangent. A standard rectified linear unit (ReLU) is an activation function that can enhance the computing performance of ANNs [53], [62]. Mathematically, ReLU is defined as

$$A = \begin{cases} X_i, & \text{if } X_i \geq 0 \\ 0, & X_i < 0 \end{cases}.$$

Deep neural networks (DNNs) have been developed to be investigated. The structure of DNN model consists of three hidden layers where each hidden layer contains 100 neurons. The activation function is ReLU function. Accordingly, DNN's structure is (4-100-100-100-1).

B. Ensemble AI models

Ensemble methods (fusion learning) are elegant data mining techniques to combine multiple learning algorithms to enhance the overall performance [37]. Ensemble methods can apply ML algorithms, such as ANN, DT, and SVM which are called "base model or base learner" as inputs for ensemble methods. The concept behind ensemble methods can be illustrated as Fig. 5 and mathematically as (7) [19]. The two questions in Fig. 5 illustrated the mechanism of tree algorithm where each question is split the main tree to these branches. The questions are for only illustration to present the decision process inside the tree model.

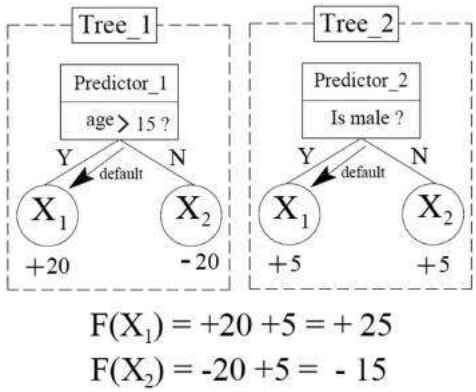

Fig. 5. Additive function concept.

For the given dataset (D) with *n* examples (144 cases) and *m* features (4 key cost drivers) $D = (x_i; y_i)$ $(x_i \in R^m; y_i \in R)$ where *R* is the real numbers set. *K* is an additive function to predict the output as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), \qquad f_k \in F \quad (7)$$

where $\boldsymbol{F = \{f(x) = w_{q(x)}\}}$ $\boldsymbol{(q: \mathbb{R}^m \rightarrow T, T \in \mathbb{R}^m)}$.

q is the structure of each tree that maps an example to the corresponding. T corresponds to the number of leaves in the tree. $\hat{y}_i$ is the predicted dependent variable (FCIPs cost LE / project). Each f_k represents an independent tree structure q and leaf weights (*w*). (X_i) represents independent variables. (*F*) represents the regression trees space.

1) Bagging: Bagging is a variance reduction algorithm to train several classifiers based on bootstrap aggregating as shown in Fig. 6(a). Bagging algorithm randomly draws replicas of training dataset with replacement to train each classifier [14], [15]. As a result, diversity is obtained by resampling several data subsets. On average, each bootstrap sample contains 63.2% of

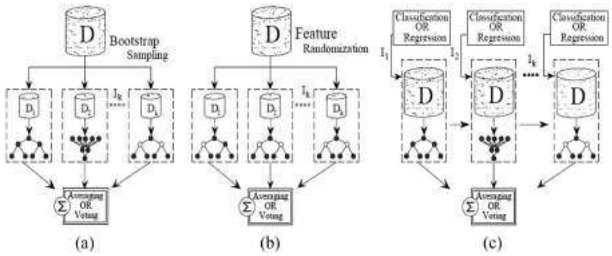

Fig. 6. (a) Bagging (b) RF, and (c) Boosting.

the original training data set. The CART is selected as a base learner for bagging model.

2) Random Forest: RF is one of bagging ensemble learning models that can produce accurate performance without over-fitting issue [16] as shown in Fig. 6(b). RF algorithm draws bootstrap samples to develop a forest of trees based on random subsets of features. Extremely randomized tree algorithm (Extra Trees) merges the randomization of random subspace to a random selection of the cut-point during splitting tree node process. Extremely randomized tree mainly controls the attribute randomization and smoothing parameters [35].

3) Boosting and Adaptive Boosting: Schapire has presented boosting procedure (also known as adaptive resampling) as an algorithm that boosts the performance of weak learning algorithms [75]. Bagging generates classifiers in parallel while boosting develops the classifiers sequentially as shown in Fig. 6(C). Thus, boosting converts weak models to strong ones. Freund and Schapire [96] have presented an Adaptive Boosting algorithm (AdaBoost). AdaBoost is selected as one of the top ten data mining algorithms. AdaBoost serially manipulates the cost data for each base learner to AdaBoost and assigns equal weights for all cases where larger weights are assigned to the misclassified cases. The objective is to make a greater focus on the misclassified cases to be corrected in the consequent iteration. In addition, the AdaBoost algorithm assigns other weights to rank each individual base learning algorithm based on its accuracy [9]. The CART is selected as a base learner for AdaBoost model to predict the conceptual cost of FCIPs.

4) Extreme Gradient Boosting: XGBoost is a large-scale ML model that can build a highly scalable end-to-end ensemble tree boosting system for big data processing [19]. XGBoost is a modified gradient tree model with regularization term to the additive function as follows:

$$L(\Phi) = (x+a)^n = \sum_{k=0}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$
$$\text{where} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{8}$$

where $\boldsymbol{L}$ represents a differentiable convex cost function that determines the difference between the predicted output $\hat{y}_i$ and the actual output y_i. Ω is a regularization term to avoid over-fitting and smooth the learned weights (W_i). The regularization term penalizes the complexity of the regression tree functions.

5) Stochastic Gradient Boosting: The performance of gradient boosting can iteratively be improved as stochastic gradient boosting (SGB) algorithm by injecting randomization into the selected data subsets. Injecting randomization to boosting algorithm can substantially boost both the fitting accuracy and computational cost of the gradient boosting algorithm ([14]; Freund and Schapire [96]). The training data are randomly drawn at each iteration without replacement from the data set. SGB can be viewed in this sense as a boosting bagging hybrid.

VI. Evaluation Techniques

Evaluation techniques for predictive models can be MAPE, the mean squared error, the root mean squared error, the coefficient of determination (R^2) or adjusted $R*^2$. MAPE is comparing the predicted and actual outcomes [59] as (3). MAPE can be classified as an excellent prediction if MAPE is less than 10%, between 10% from 20% is a good prediction. Between 20% and 50% is acceptable forecasting and more than 50% is an inaccurate prediction [55]. Based on Peurifoy and Oberlender [68], this study has categorized model accuracy to three main categories

$$MAPE\ \%\ categorization = \begin{cases} below\ 10, & 10\% \geq MAPE \geq 0 \\ below\ 20, & 20\% \geq MAPE > 10\% \\ unacceptable, & MAPE > 20\% \end{cases}$$

where "below 10" indicates a high accuracy level than "below 20".

The R-squared (R^2: coefficient of determination) is expressed as follows:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$
$$-1 \leq R^2 \leq 1 \tag{9}$$

where SSE is the sum of squares of the residuals and SST is the total sum of squares. $\bar{y}_i$ is the arithmetic mean of the Y variable. R^2 measures the percentage of the variation of the predictor y_i explained by the dependent variable X. Thus, R^2 indicates how well the model fits the cost data. IF R^2 value is 0.9 or above, it is classified as very good, above 0.8 is good, above 0.5 is satisfactory, and below 0.5 is poor [3], [64]. Adjusted R-squared R^{*2} is computed as follows:

$$R^{*2} = R^2 - \frac{(1 - R^2) K}{n - (K+1)} \tag{10}$$

where R^{*2} is adjusted for the number of variables included in the proposed model where R^{*2} is lower than R^2 value. For model evaluation, R^{*2} is always preferred to R^2 to avoid the over-fitting problem [3], [64].

VII. Results, Comparison, and Analysis

MAPE and R^{*2} have been validated the 20 developed models as displayed in Table III. The whole developed models have been descendingly sorted from M 1 to M 20 based on MAPE as shown in Fig. 7.

TABLE III
ACCURACY OF THE DEVELOPED ALGORITHMS

Notation	Algorithm / model	Algorithm type: supervised regression	MAPE %	MAPE % categorization	R^2	R^{*2}
M 1	XGBoost	Ensemble methods	9.091	below 10	0.931	0.929
M 2	Quadratic regression*	MRA	9.120	below 10	0.857	0.851
M 3	Plain regression*	MRA	9.130	below 10	0.803	0.796
M 4	Quadratic MLP*	ANNs	9.200	below 10	0.904	0.902
M 5	Plain MLP*	ANNs	9.270	below 10	0.913	0.912
M 6	Semilog regression*	MRA	9.300	below 10	0.915	0.910
M 7	Extra Trees	Ensemble methods	9.714	below 10	0.948	0.947
M 8	Natural log MLP*	ANNs	10.230	below 20	0.905	0.910
M 9	Bagging	Ensemble methods	10.246	below 20	0.914	0.911
M 10	RF	Ensemble methods	10.503	below 20	0.916	0.913
M 11	AdaBoost	Ensemble methods	10.679	below 20	0.875	0.871
M 12	SGB	Ensemble methods	11.008	below 20	0.926	0.924
M 13	Reciprocal regression*	MRA	11.200	below 20	0.814	0.801
M 14	Power (2) regression*	MRA	11.790	below 20	0.937	0.931
M 15	DNNs	ANNs	12.059	below 20	0.785	0.779
M 16	DT	Tree model	12.488	below 20	0.886	0.883
M 17	Genetic Fuzzy	Hybrid model	**12.900**	below 20	**0.893**	**0.890**
M 18	CBR	Case based	17.300	below 20	0.859	0.852
M 19	SVM	Kernel based	21.217	unacceptable	0.136	0.133
M 20	Fuzzy	Fuzzy theory	26.300	unacceptable	0.857	0.851

***: (Elmousalami et at. 2018 b)**

Based on Table III, Fig. 8 shows that ensemble methods formulate six out of 20 models and MRA and ANNs formulates five and four models, respectively. As a result, ensemble methods can generate infinite models using different base learners. Similarly, MRA and ANNs can develop several models using data transformation and different training weights. Out of 20 the developed models, only 35% produced the accuracy below 10 of MAPE. This is the key insight of comparison of different ML algorithms where 10% produced unacceptable performance and a total of 55% produced MAPE more than 10%. As a result, this analysis highlights the performance of ensemble methods, MRA, and ANNs for high accurate cost prediction.

Elmousalami *et al*. [31] have presented quadratic regression model (M 2) as the most accurate for FCIPs among the developed regression and ANNs models (M3, M4, M5, M6, M8, M13, and M14) with 9.120 and 0.851 for MAPE and R^{*2}, respectively. However, this study presents that XGBoost (M 1) is more accurate than quadratic regression (M 2). XGBoost (M 1) have obtained the first place slightly higher than M 2 with 9.091% and 0.929 for MAPE and R^{*2}, respectively. Moreover, the unique advantage of the XGBoost is its high scalability where it can process noisy data and fit high dimension data without overfitting. XGBoost applies parallel computing to effectively reduce computational complexity and learn faster [19]. Another key advantage of XGBoost is handling the missing values where defaults direction is identified as shown in Fig. 5. Accordingly, no effort is needed for cleaning the collected data.

Ensemble methods such as [Extra Trees (M 7), bagging (M 9), RF (M 10), AdaBoost (M 11), and SGB (M 12)] have produced a high acceptable performance where its accuracy is ranging from 9.714 to 11.008%. The ensemble learning methods can effectively deal with the problems of high-dimension data, complex data structures, and small sample size. Bagging algorithms can increase generalization by decreasing variance error (Breiman [97]) while boosting can improve generalization by decreasing bias error [76]. Ensemble methods can effectively handle continues, categorical, and dummy features with missing values. However, ensemble methods increase the model complexity which decreases the model interpretability [52]. RF (M 10) is a robust algorithm against noisy data or big data than the DT (M 16) algorithm [14], [26]. However, the RF algorithm is unable to interpret the importance of features or the mechanism of producing the results.

DNNs (M 15) produce 12.059% MAPE less than all the developed MLP (M 4, M 5, and M 8). Accordingly, DNNs provide bad performances with a small dataset. Conversely, deep learning and DNNs can produce the most accurate performance with high dimension data [53]. An alternative to the black box

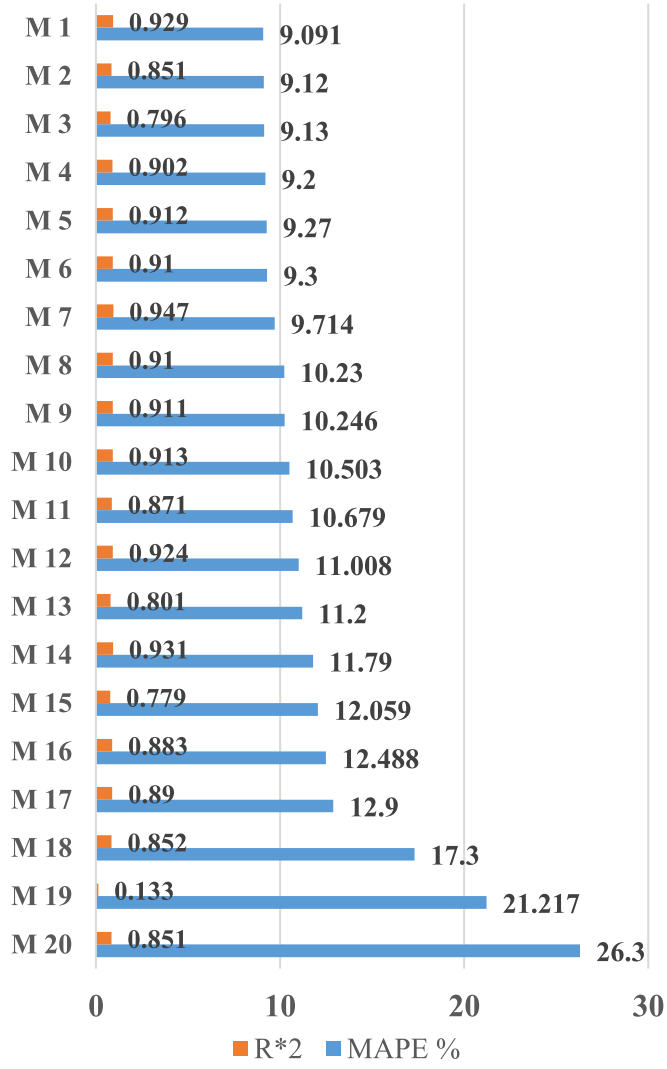

Fig. 7. MAPE and R^{*2} for all algorithms.

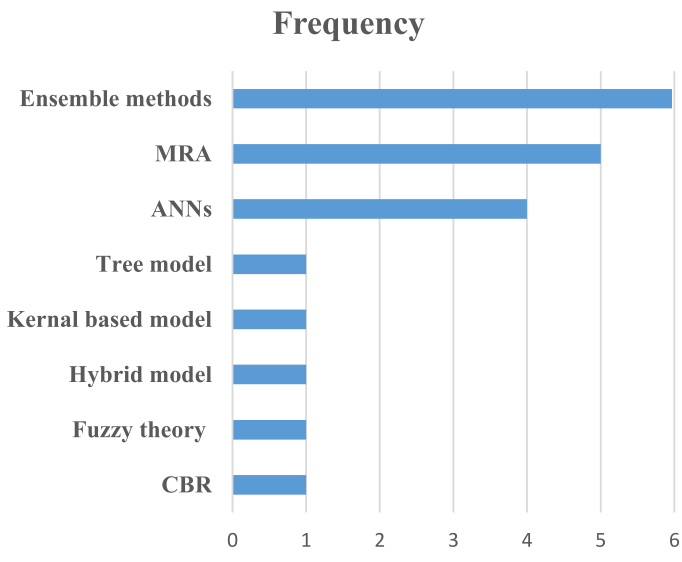

Fig. 8. Frequency of each model category.

nature of ANNs and DNNs, DT generates logic statements and interpretable rules which can be used for identifying the importance of data features [67]. Another advantage of DT is avoiding the curse dimensionality and providing a high-performance computing efficiency through its splitting procedure [70]. However, DT is producing unsatisfactory performance in time series, noisy, or nonlinear data [25]. Although DT (CART) is inherently used as a based learner for the ensemble methods, DT (M 16) produces 12.488% MAPE less than all developed ensemble methods (M1, M7, M9, M10, M11, and M12).

Tree models can provide high performance and predictive accuracy using small data set because tree and ensemble algorithms classify the small data using pivot questions to form the tree branches. Conversely, regression, SVM, ANNs and deep learning needs more data to train the neurons weights and minimize the predictive error. Tree models produces pivot decision making on each node to classify the data; these questions or decision interpret the mechanism that the tree conducts to predict the output. On the other hand, SVM, ANNs, and deep learning are Blackbox algorithms. Tree models-based models such as XGBoost and RFs have defaults direction that guide the model to handle the missing data. On the other hand, SVM, ANNs, and deep learning cannot process the missing values and the developer must clean this missing data as preprocessing stage before using ANNs or regression algorithms.

Fuzzy models can maintain uncertainty to cost estimate where Tree models, regression and ANNs are deterministic values producers. Ensemble models are modeling techniques to boost the performance of the single algorithms such as tree models, regression and ANNs.

Therefore, ensemble models for tree algorithms (Tree model as a base learner) such as random forest or XGBoost model will have the advantage of the single tree model such as handling small data size, missing data values, and interpretation. Moreover, the performance of the ensemble of tree algorithms will produce more accurate performance than single tree. Therefore, XGBoost produced the optimal performance among all the 20 algorithms as shown in Table II.

Ensemble methods and data transformation play an important role in prediction accuracy. However, the main gap of the previous models is lacking uncertainty modeling to the prediction cost model. Therefore, FL theory has been conducted to maintain uncertainty concept through the FL model (M 17) and hybrid fuzzy model (M 20). The number of generated rules by the fuzzy genetic model (M 17) is 63 and the MAPE is 14.7%. On the other hand, a traditional FL model (M 20) has been built based on the experts 'experience where a total of 190 rules are generated to cover all the possible combinations of the fuzzy system and MAPE is 26.3%. Moreover, the fuzzy rules (IF–Then rules) generated by experts have redundant rules which can be deleted to improve the model computation and performance. Moreover, the expert's knowledge cannot cover all combination to represent all possible rules (2401 rules). In addition, the generation of the experts' rules is time and effort consuming process. Consequently, hybrid fuzzy systems are more effective than the traditional FL system. Although the

TABLE IV
CHARACTERISTICS OF THE DEVELOPED ALGORITHMS

	Strengths	Weaknesses	nterpretation	Uncertainty	Missing values
M 1	High scalability, handing missing values, high accuracy, low computational cost	No uncertainty and interpretation	no	no	yes
M 2	More accurate then plain regression, hand nonlinearity of data	Prone to overfitting	yes	no	no
M 3	Works on small size of dataset	Linear assumptions	yes	no	no
M 4	High accuracy, handling complex patterns	Black box nature, need sufficient data for training	no	no	no
M 5	High accuracy, handling complex patterns	Black box nature	no	no	no
M 6	Producing better results than plain regression	Unable to capture complex patterns	yes	no	no
M 7	Handing data randomness	Black box nature and sufficient data	no	no	yes
M 8	Producing better results than plain MLP	Black box nature and sufficient data	no	no	no
M 9	Providing higher performance than a single algorithm	Depending on other algorithms performance	no	no	yes
M 10	Accurate and high performance on many problems including non linear	No interpretability, need to choose the number of trees	no	no	yes
M 11	High scalability, and high adaptability	Depends on other algorithms performance	no	no	yes
M 12	Handing difficult examples	Highly sensitive to noisy data	no	no	yes
M 13	Handing data nonlinearity and training small sample size	Unable to capture complex patterns	yes	no	no
M 14	Handing data nonlinearity and training small sample size	Unable capture complex patterns	yes	no	no
M 15	Capturing complex patterns, processing big data and high-performance computing	Sufficient training data and high cost computation	no	no	no
M 16	Working on both linear and nonlinear data, and producing logical expressions	Poor results on too small datasets, overfitting can easily occur	yes	no	no
M 17	Handing uncertainty and more accurate than fuzzy model	More complex that fuzzy model and needs more computational resources	yes	yes	no
M 18	Handling small data sets, simple and needs less computational time	Poor performance where the optimal case cannot be retrieved	yes	no	no
M 19	Easily adaptable, works very well on nonlinear problems, not biased by outliers	Compulsory to apply feature scaling, more difficult to understand	no	no	no
M 20	Handing uncertainty	Low accuracy	yes	yes	no

prediction accuracy of the fuzzy genetic model (M 17) and the FL model (M 20) is 14.7% and 26.3%, respectively, the fuzzy model would produce more reliable prediction results by taking uncertainty into account. However, the traditional fuzzy model gives unacceptable accuracy of 26.3% MAPE [68]. Therefore, maintaining uncertainty decreases predictive model accuracy.

CBR (M-18) produces an acceptable low accuracy of 17.3% MAPE. The advantage of CBR is dealing with a vast amount of data where all past cases and new cases are stored in database techniques [47]. Moreover, finding similarities and similar cases improve the reliability and confidence in the output. Hybrid models can be incorporated to CBR to enhance the performance of CBR such as applying GA and DT to optimize attributes weights and applying regression analysis for the revising process. SVM can be applied for both regression and classification tasks. SVM (M 19) produces unacceptable accuracy of 21.217% MAPE [68]. Finally, Table IV summarizes the strengths and weakness of each developed model.

VIII. Conclusion

This article presented a comparison of AI techniques to develop a reliable conceptual cost prediction model. A total of 20 ML models were developed utilizing tree-based models, ensemble methods, fuzzy systems, CBR, ANNs, SVM, and transformed regression models. The accuracy of the developed models was tested from two perspectives: MAPE and adjusted R2. The results showed that the most accurate and suitable method was XGBoost with 9.091% and 0.929% for MAPE and adjusted R2, respectively. The study emphasized the importance of ensemble methods for improving the prediction accuracy, and handling noisy and missing data. However, the key limitation of the ensemble methods was an inability to interpret the producing results. In addition, DT algorithm and ensemble methods could provide an alternative technique to many ML algorithms such as MRA and ANNs.

The conceptual cost estimate was conducted under uncertainty. Therefore, this study recommended using fuzzy theory such as FL and to develop a hybrid model based on FL to obtain uncertainty nature for the developed model and produce more reliable performance. In addition, the study highlighted the main challenge for fuzzy modeling which was fuzzy inference rules generation. This study had discussed the importance of the hybrid fuzzy model methodologies to generate rules such as genetic fuzzy model. Therefore, this study recommended developing an automated hybrid fuzzy rules models than traditional fuzzy models. The fusion of the AI techniques was called hybrid intelligent systems where Zadeh [89] had predicted that the hybrid intelligent systems will be the way of the future.

The key contribution and insights could be summarized as follows.

1) Elmousalami *et al*. [31] claimed that the quadratic regression model was the optimal model for conceptual cost prediction using FCIPs data. However, this study showed that XGBoost model can present better accuracy the quadratic model.
2) This study presented a comprehensive guide to develop an automated model for the parametric cost model or conceptual cost model at the early stage of the project.
3) This study presented the first application of XGBoost algorithm for conceptual cost estimate.
4) This study addressed the ensemble ML algorithms for cost estimates. Ensemble ML algorithms produces a superior performance than single ML.
5) Out of 20 the developed models, only 35% produced the accuracy below 10 of MAPE. This was the key insight of comparison of different ML algorithms where 10% produced unacceptable performance and a total of 55% produced MAPE more than 10%. As a result, this analysis highlighted the performance of ensemble methods, MRA, and ANNs for high accurate cost prediction as shown in Fig. 9.

Future research can be stated as follows.

1) Deep learning is a powerful tool for pattern recognition. Therefore, the future trend of the cost estimates may rely on deep neural networks based on the big data of the construction projects.
2) The hybrid model represents the current trend of parametric cost modeling to improve the model performance and accuracy. For example, hybrid models can be incorporated to CBR to enhance the performance of CBR such as by applying GA and DT to optimize attribute weights and by applying regression analysis for the revision process.
3) There is a need to develop a model that has the ability to give justification for the model's results and to give answers and interpretations for the predicted cost. That may require a higher level of AI and may represent the future trend of cost modeling. Moreover, such concept may be generalized for any prediction model. The objective is to avoid the estimator's biases, warn the user to the input parameters of the model, and avoid the limitation of the black box nature.
4) The conceptual cost estimate is conducted under uncertainty. Therefore, this study recommends using fuzzy theory such as FL and to develop a hybrid model based on FL to obtain uncertainty for the developed model and produce a more reliable performance [29].
5) Ensemble methods are promising techniques that can handle a large number of features, model both numerical and categorical variables, capture nonlinear patterns, and fit data with missing values.
6) DT algorithms and ensemble methods can provide an alternative technique to many ML algorithms such as MRA and ANNs. The study emphasizes the importance of ensemble methods for improving the prediction accuracy and handling noisy and missing data. However, the key limitation of the ensemble methods in an inability to interpret the produced results.[1]

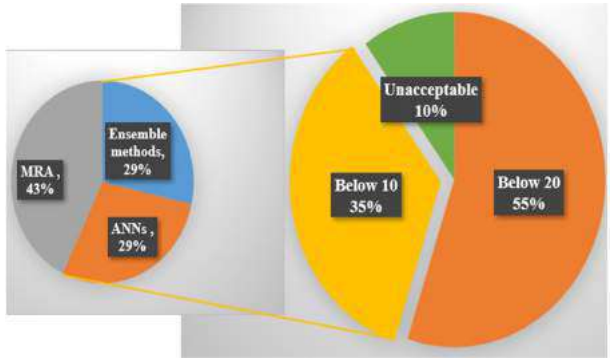

Fig. 9. Models accuracy categorization.

References

[1] *AACE International Recommended Practices*, AACE International, 2004.
[2] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, no. 1, pp. 39–59, 1994.

[1]Dataset available at https://github.com/HaythamElmousalami/Field-canals-improvement-projects-FCIPs-

[3] A. D. Aczel, *Complete Business Statistics*. Homewood, IL, USA: Irwin, 1989, p. 1056.

[4] A. Alroomi, D. H. S. Jeong, and G. D. Oberlender, "Analysis of cost-estimating competencies using criticality matrix and factor analysis," *J. Construction Eng. Manage.*, vol. 138, no. 11, pp. 1270–1280, 2012.

[5] A. C. Amason, "Distinguishing the effects of functional and dysfunctional conflict on strategic decision making: Resolving a paradox for top management teams," *Acad. Manage. J.*, vol. 39, no. 1, pp. 123–148, 1996.

[6] V. R. Ambrule and A. N. Bhirud, "Use of artificial neural network for pre design cost estimation of building projects," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 5, no. 2, pp. 173–176, 2017.

[7] P. P. Angelov, "Evolving rule-based models," *A Tool for Design of Flexible Adaptive Systems*, Wurzburg, Germany: Physica-Verlag, 2002.

[8] V. Arabzadeh, S. T. A. Niaki, and V. Arabzadeh, "Construction cost estimation of spherical storage tanks: artificial neural networks and hybrid regression—GA algorithms," *J. Ind. Eng. Int.*, vol. 14, pp. 747–756, 2018.

[9] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Mach. Learn.*, vol. 36, no. 1/2, pp. 105–139, 1999.

[10] M. J. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Hoboken, NJ, USA: Wiley, 1997.

[11] J. C. Bezdek, "What is computational intelligence?" in *Computational Intelligence Imitating Life*. J. M. Zurada, R. J. Marks II, and C. J. Robinson, Eds, New York, NY, USA: IEEE Press, 1994, pp. 1–12.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer, pp. 1–58, 2006.

[13] J. Bode, "Neural networks for cost estimation: simulations and pilot application," *Int. J. Prod. Res.*, vol. 38, no. 6, pp. 1231–1254, 2000.

[14] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, pp. 123–140, 1996.

[15] L. Breiman, "Pasting small votes for classification in large databases and on-line," *Mach. Learn.*, vol. 36, no. 1/2, pp. 85–103, 1999.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[17] L. Breiman, J. H. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[18] Y. Cao, B. Ashuri, and M. Baek, "Prediction of unit price bids of resurfacing highway projects through ensemble machine learning," *J. Comput. Civil Eng.*, vol. 32, no. 5, 2018, Art no. 04018043.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[20] M.-Y. Cheng and A. F. Roy, "Evolutionary fuzzy decision model for construction management using support vector machine," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 6061–6069, 2010.

[21] M.-Y. Cheng, H.-C. Tsai, and W.-S. Hsieh, "Web-based conceptual cost estimates for construction projects using evolutionary fuzzy neural inference model." *Autom. Construction*, vol. 18, no. 2, pp. 164–172, 2009.

[22] C.-H. Chou, "Genetic algorithm-based optimal fuzzy controller design in the linguistic space," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 3, pp. 372–385, Jun. 2006.

[23] J. S. Chou and C. Lin, "Predicting disputes in public-private partnership projects: Classification and ensemble models," *J. Comput. Civil Eng.*, vol. 27, no. 1, pp. 51–60, 2012.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[25] S. P. Curram and J. Mingers, "Neural networks, decision tree induction and discriminant analysis: An empirical comparison," *J. Oper. Res. Soc.*, vol. 45, no. 4, pp. 440–450, 1994.

[26] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[27] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton, "JPred4: A protein secondary structure prediction server," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W389–W394, 2015.

[28] O. Dursun and C. Stoy, "Conceptual estimation of construction costs using the multistep ahead approach," *J. Construction Eng. Manage.*, vol. 142, no. 9, 2016, Art no. 04016038.

[29] H. H. Elmousalami, "Intelligent methodology for project conceptual cost prediction," *Heliyon*, vol. 5, no. 5, 2019, Art no. e01625.

[30] H. H. Elmousalami, A. H. Elyamany, and A. H. Ibrahim, "Evaluation of cost drivers for field canals improvement projects," *Water Resour. Manage.*, vol. 32, no. 1, pp. 53–65, 2018.

[31] H. H. Elmousalami, A. H. Elyamany, and A. H. Ibrahim, "Predicting conceptual cost for field canal improvement projects," *J. Construction Eng. Manage.*, vol. 144, no. 11, 2018, Art no. 04018102.

[32] G. Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: A decision tree approach," *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, 2006.

[33] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[34] W. - F. Feng, W. - J. Zhu, and Y. - G. Zhou, "The application of genetic fuzzy neural network in project cost estimate," in *Proc. Int. Conf. E-Product E-Service E-Entertainment*, 2010, pp. 1–4.

[35] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[36] S. B. Green, "How many subjects does it take to do a regression analysis?" *Multivariate Behav. Res.*, vol. 26, pp. 499–510, 1991.

[37] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[38] T. Hatanaka, Y. Kawaguchi, and K. Uosaki, "Nonlinear system identification based on evolutionary fuzzy modelling," in *Proc. IEEE Congr. Evol. Comput.*, 2004, vol. 1, 646–651.

[39] T. Hegazy and A. Ayed, "Neural network model for parametric cost estimation of highway projects," *J. Construction Eng. Manage.*, vol. 124, no. 3, pp. 210–218, 1998.

[40] M. Ilbeigi, B.Ashuri, and A. Joukar, "Time-series analysis for forecasting asphalt-cement price," *J. Manage. Eng.*, vol. 33, no. 1, 2016, Art no. 04016030.

[41] R. Jin, K. Cho, C. Hyun, and M. Son, "MRA-based revised CBR model for cost prediction in the early stage of construction project," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5214–5222, 2012.

[42] A. Jrade and S. Alkass, "A conceptual cost estimating computer system for building projects," presented at the *Proc. AACE Int. Trans.*, 2001, Paper IT91.

[43] M. Juszczyk, "Studies on the ANN implementation in the macro BIM cost analyzes," *Przegląd Naukowy. Inżynieria i Kształtowanie Środowiska*, vol. 26, no. 2, pp. 183–192, 2017.

[44] M. Juszczyk, A. Leśniak, and K. Zima, "ANN based approach for estimation of construction costs of sports fields," *Complexity*, vol. 26, 2018, Art. no. 7952434.

[45] Y. Karatas and F. Ince, "Feature article: Fuzzy expert tool for small satellite cost estimation," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 31, no. 5, pp. 28–35, May 2016.

[46] G. -H. Kim, S. -H. An, and K. -I. Kang, "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning," *Building Environ.*, vol. 39, no. 10, pp. 1235–1242, 2004.

[47] S. Kim, "Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation," *J. Civil Eng. Manage.*, vol. 19, no. 1, pp. 86–96, 2013.

[48] S. Kim, S. Chin, and S. Kwon, "A discrepancy analysis of BIM-based quantity take-off for building interior components," *J. Manage. Eng.*, vol. 35, no. 3, 2019, Art no. 05019001.

[49] B. Kosko and S. Mitaim, "What is the best shape for a fuzzy set in function approximation?" in *Proc. 5th IEEE Int. Conf. Fuzzy Syst.*, 1996 pp. 1237–1243.

[50] J. L. Kolodner, "An introduction to case-based reasoning," *Artif. Intell. Rev.*, vol. 6, pp. 3–34, 1992.

[51] L. I. Kuncheva, *Combining Pattern Classifiers: Methods And Algorithms*. Hoboken, NJ, USA: Wiley, 2004.

[52] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[53] A. Leśniak and K. Zima, "Cost calculation of construction projects including sustainability factors using the Case Based Reasoning (CBR) method," *Sustainability*, vol. 10, 5, 2018, Art no. 1608.

[54] C. D. Lewis, *Industrial and Business Forecasting Methods*. London, U.K.: Butterworths, 1982.

[55] B. P. Loop, S. D. Sudhoff, S. H. Zak, and E. L. Zivi, "Estimating regions of asymptotic stability of power electronics systems using genetic algorithms," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 5, pp. 1011–1022, Sep. 2010.

[56] D. J. Lowe, M. W. Emsley, and A. Harding, "Predicting construction cost using multiple regression techniques," *J. Construction Eng. Manage.*, vol. 132, no. 7, pp. 750–758, 2006.

[57] M. D. Dell'Isola, *Architect's Essentials of Cost Management*. New York, NY, USA: Wiley, 2002.

[58] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting Methods and Applications*. New York, NY, USA: Wiley, 1998.

[59] M. Marzouk and M. Alaraby, "Predicting telecommunication tower costs using fuzzy subtractive clustering," *J. Civil Eng. Manage.*, vol. 21, no. 1, pp. 67–74, 2014.

[60] M. Marzouk and M. Elkadi, "Estimating water treatment plants costs using factor analysis and artificial neural networks," *J. Cleaner Prod.*, vol. 112, pp. 4540–4549, 2016.
[61] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[62] A. J. Ogungbile, A. E. Oke, and K. Rasak, "Developing cost model for preliminary estimate of road projects in Nigeria," *Int. J. Sustain. Real Estate Construction Econ.*, vol. 1, no. 2, pp. 182–199, 2018.
[63] E. Ostertagová, *Applied Statistic*. Elfa Košice, Slovakia: Tech. Univ. Košice (in Slovak), 2011, p. 161.
[64] W. Pedrycz Ed., *Fuzzy Evolutionary Computation*. Dordrecht, The Netherlands: Kluwer, 1997, vol. 1197.
[65] S. Perera and I. Watson, "Collaborative case-based estimating and design," *Advances Eng. Softw.*, vol. 29, no. 10, pp. 801–808, 1998.
[66] P. Perner, U. Zscherpel, and C. Jacobsen, "A comparison between neural networks and decision trees based on data from industrial radiographic testing," *Pattern Recognit. Lett.*, vol. 22, no. 1, pp. 47–54, 2001.
[67] R. L. Peurifoy and G. D. Oberlender, *Estimating Construction Costs*, 5th ed. New York, NY, USA: McGraw-Hill, 2002.
[68] A. PMI, *Guide to the Project Management Body of Knowledge: PMBOK 2000*. Newtown Square, PA, USA: Project Manage. Inst., 2008.
[69] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.
[70] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[71] H. G. Radwan, "Sensitivity analysis of head loss equations on the design of improved irrigation on-farm system in Egypt," *Int. J. Adv. Res. Technol.*, vol. 2, no. 1, 2013.
[72] B. H. Ross, "Some psychological results on case-based reasoning," in *Proc. Case-Based Reason. Workshop*, Burlington, MA, USA, 1989, pp. 144–147.
[73] L. Sabol, "Challenges in cost estimating with building information modeling," *IFMA World Workplace*, pp. 1–16, 2008.
[74] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.
[75] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.
[76] A. Shreenaath, S. Arunmozhi, and R. Sivagamasundari, "Prediction of construction cost overrun in Tamil Nadu—-A statistical fuzzy approach," *Int. J. Eng. Techn. Res.*, vol. 3, no. 3, pp. 267—275, 2015.
[77] N. Siddique and H. Adeli, *Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing*. Chichester, U.K.: Wiley, 2013.
[78] C. Stoy, S. Pollalis, and O. Dursun, "A concept for developing construction element cost models for German residential building projects," *Int. J. Project Organisation Manage.*, vol. 4, no. 1, pp. 38–53, 2012.
[79] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. Moscow, Russia: Nauka (in Russian), 1979, pp. 5165–5184.
[80] Y. -R. Wang, C. -Y. Yu, and H. -H. Chan, "Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models," *Int. J. Project Manage.*, vol. 30, no. 4, pp. 470–478, 2012.
[81] T. P. Williams and J. Gong, "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers," *Autom. Construction*, vol. 43, pp. 23–29, 2014.
[82] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.
[83] A. Woldesenbet and D. H. S., Jeong, "Historical data driven and component based prediction models for predicting preliminary engineering costs of roadway projects," *Proc. Construction Res. Congr.*, 2012.
[84] I.-T. Yang, "Simulation-based estimation for correlated cost elements," *Int. J. Project Manage.*, vol. 23, no. 4, pp. 275–282, 2005.
[85] F. Yoav and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, 1999, pp. 148–156.
[86] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.
[87] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision process," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-3, no. 1, pp. 28–44, Jan. 1973.
[88] L. A. Zadeh, "Fuzzy logic, neural networks and soft computing," *Commun. ACM*, vol. 37, pp. 77–84, 1994.
[89] K. Zhai, N. Jiang, and W. Pedrycz, "Cost prediction method based on an improved fuzzy model," *Int. J. Adv. Manuf. Technol.*, vol. 65, no. 5/8, pp. 1045–1053, 2012.
[90] R. Zhang, B. Ashuri, Y. Shyr, and Y. Deng, "Forecasting construction cost index based on visibility graph: A network approach," *Physica A: Statistical Mech. Appl.*, vol. 493, pp. 239–252, 2018.
[91] W. Zhang and Y. Liu, "Fuzzy logic controlled particle swarm for reactive power optimization considering voltage stability," in *Proc. IEEE Int. Conf. Power Eng.*, Singapore, 2005, pp. 1–5.
[92] J. Wang and B. Ashuri, "Predicting ENR construction cost index using machine-learning algorithms," *Int. J. Construction Edu. Res.*, vol. 13, no. 1, pp. 47–63, Jan. 2017.
[93] M, Ranasinghe, "Impact of correlation and induced correlation on the estimation of project cost of buildings," *Construction Manag. Econ.*, vol. 18, no. 4, pp. 395–406, Jun. 2000.
[94] C. Stoy, S. Pollalis, and H. R. Schalcher, "Drivers for cost estimating in early design: Case study of residential construction," *J. Construction Eng. Manag.*, vol. 134, no. 1, pp. 32–39, Jan. 2008.
[95] W. J. Zhu, W. F. Feng, and Y. G. Zhou, "The application of genetic fuzzy neural network in project cost estimate," in *Proc. Int. Conf. E-Prod. E-Serv., E-Entertain.*, Nov. 7, 2010, pp. 1–4.
[96] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.
[97] L. Breiman, "Arcing classifier (with discussion and a rejoinder by the author)," *Annals Statist.*, vol. 26, no. 3, pp. 801–849, 1998.