Feature Selection for Construction Organizational Competencies Impacting Performance

Getaneh Gezahegne Tiruneh *PhD Candidate* [1] and Aminah Robinson Fayek *Professor*
[1]*Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada*

Abstract**—**Organizational competencies have a significant influence on performance; therefore, it is vital that organizations in the construction industry assess and enhance their competencies in order to improve performance. The set of variables that captures construction organizational competencies is highly dimensional. Feature selection (FS) helps to reduce the dimensionality of data by using only a subset of variables to develop a model. The main objective of the research presented in this paper is the development of a fuzzy inference system (FIS) applying fuzzy c-means clustering (FCM) and FS using genetic algorithms (GAs). First, the parameters of FCM are optimized and used to develop an FIS. Then, the FIS is optimized using a GA. The root mean square error (RMSE) is used as a fitness function for GA optimization. This paper contributes an FS approach for construction engineering and management problems that are characterized by high dimensionality of feature space and few data instances.

Keywords**—***feature selection, fuzzy c-means clustering, genetic algorithm, competency, performance*

I. Introduction

The environment within which the construction industry operates is dynamic and complex in nature, leading to the increasing presence of uncertainties in technology, budgets, and development processes [1]. The construction industry demands continuous quality, productivity, and performance improvement, which is attributable to the emergence of new procurement methods, contracts, and project delivery methods [2]. However, the construction industry is criticized for its underperformance and continues to suffer from declining productivity [2], [3]. Chakravarty et al. in [4] suggest that in order to have an impact, multiple competencies must be directed towards the implementation of competitive actions that will enhance an organization's performance. Evolution in theory and practice has placed competencies at the center of an organization's success [5]. Therefore, in order to achieve better performance and competitiveness, it is vital for construction organizations to explore new approaches for assessing and enhancing their competencies [6], [7]. Tiruneh and Fayek define organizational competency in [8], [9] as "an integrated combination of resources, particular sets of skills, necessary information, technologies, and the right corporate culture that enable an organization to achieve its corporate goals, competitive advantage, and superior performance." Relating organizational competencies to performance is essential for identifying target areas to improve performance.

The variables that characterize construction organizational competencies are both quantitative and qualitative, and in most cases, the set of variables that captures organizational competencies is highly dimensional, which leads to long data processing time and low accuracy of the predictive model for an organization's performance. Processing the data to obtain a smaller set of representative features while retaining optimal salient characteristics of the data not only decreases the processing time, but also leads to more compactness of the models and better generalization [10]. A suitable way to overcome this problem (i.e., the high dimensionality of the feature space) is to implement dimensionality reduction, which reduces the dimensionality of the data by using only a subset of variables to develop a model [11]. Dimensionality reduction makes it possible to obtain a smaller set of representative organizational competency variables that retains the optimal relevant characteristics of the data, not only decreasing the processing time, but also leading to a more accurate and concise model and better generalization [11]–[15]. In general, two categories of dimensionality reduction have been commonly performed: feature extraction (transforming the existing features into a new reduced set of features) and feature selection (FS; selecting a subset of existing features) [14], [16]–[19]. This paper uses FS instead of feature extraction for selecting the best feature subsets that represent the original data.

II. Feature Selection

In data with high dimensionality, finding the optimal feature subset is a difficult task [12]. FS is the process of obtaining optimal or relevant features or a candidate subset of features from the original input features [13]–[15], [18]–[21]. The optimality of a feature subset is measured by an evaluation criterion [12], [14]. FS is an important and frequently used technique for reducing the number of features by removing irrelevant, redundant, or noisy data. FS has an immediate effect on applications, speeding up data mining algorithms and improving mining performance (e.g., predictive accuracy and result comprehensibility) [13], [19], [21]. There are three general approaches for FS: filter, wrappers, and embedded methods [14], [16], [21].

A. Feature selection methods

The filter method uses an independent measure to assess the importance of features statistically according to a heuristic criterion (i.e., a ‘relevance index’ or ‘scoring’) without involving a learning algorithm [15], [20], [21]. However, filter methods can miss features that are not useful by themselves, but that can be very useful when combined with others [12]; hence, the selected subset might not be optimal [20]. The wrapper method uses a learning algorithm (i.e., machine learning) for subset evaluation and selects an optimal subset that is best suited to a learning algorithm [12], [18]. Embedded methods combine and utilize the qualities of the filter and wrapper methods for data with high dimensionality [12], [18]. In general, embedded methods use the filter method (i.e., independent criteria) to decide the optimal subsets for pre-selection, and then the wrapper method (i.e., learning algorithm) is used to select the final optimal subset from among the optimal subsets to get high accuracy [12], [18], [21].

B. Comparison of feature selection methods

Each FS method has its own advantages and disadvantages [21]. A good FS method should have high learning accuracy and computational efficiency [18], [19]. A comparison of FS methods is presented in Table I. Wrapper methods have high learning capacity; hence, they usually obtain higher accuracy than embedded methods, which in turn are better than filter methods. On the other hand, filters are the fastest among all the methods as they need not incorporate learning, while wrappers are the slowest since they typically need to evaluate the candidate optimal feature subsets at each iteration [19], [21]. Of the three types of method discussed in this paper, wrapper methods have the highest computational complexity, while filter methods have the lowest computational complexity [12]. In dealing with extremely dimensional data for a particular application, various FS algorithms can be applied and the one that best meets the required criteria can be selected [13], [20].

TABLE I. COMPARISON OF FS METHODS

FS method	**Criteria**			
	Evaluation method	***Performance (accuracy)***	***Computational speed***	***Computational complexity***
Filter	Independent of learning algorithm	Lower than wrapper and embedded	Faster than wrapper and embedded	Lower than wrapper and embedded
Wrapper	Dependent on learning algorithm	Higher than filter and embedded	Slower than filter and embedded	Higher than filter and embedded
Embedded	Hybrid of filter and wrapper	Lower than wrapper and higher than filter	Faster than wrapper and slower than filter	Lower than wrapper and higher than filter

III. EXPERIMENTAL CASE STUDY

The experiment demonstrates the procedures required for carrying out FS for organizational competencies. First, the data is prepared. Then, an FIS is developed by conducting FCM. Finally, FS is performed using GA. A detailed description of each of the steps is presented in subsequent subsections. The steps followed in the experiment can be used to conduct FS for similar research problems in construction management and engineering.

A. Feature selection for organizational competencies impacting performance

FS techniques are used in only a few studies in the construction engineering and management domain. For example, Liu and El-Gohary in [15] used feature discretization and selection methods to evaluate the performance of data-driven bridge deterioration prediction. Tsehayae and Fayek in [22] used a filter-based FS algorithm to identify key influencing parameters for construction labor productivity. Research in the area of construction engineering and management is characterized by high dimensionality of features (parameters or variables) compared to their associated data instances. As the dimensionality of features greatly exceeds the number of data instances, it results in over fitting [16], [17], [23]. Identifying organizational competencies that impact performance shares similar challenges. From an engineering point of view, data are best characterized using as few variables as possible [23]. Therefore, an FS approach that is suitable for high dimensionality of features and limited data instances is critical for obtaining features that represent the original feature subset well. FS using population-based or evolutionary algorithms such as a genetic algorithm (GA), ant colony optimization, and particle swarm optimization are employed, which can yield optimum results and are computationally feasible [11], [19]–[21]. This paper uses GA optimization for selecting organizational competencies impacting performance. The procedure for conducting FS selection on organizational competencies that impact performance is shown in Fig. 1. The FS selection encompasses three major steps: data preparation, developing the fuzzy inference system (FIS) using fuzzy c-means clustering (FCM), and conducting FS using GA. Each of the steps shown in Fig. 1 are discussed below.

B. Data preparation

Organizational competencies are grouped as technical (how the organization operates and functions) and behavioral (individual attributes; i.e., knowledge, skills, and abilities) [24]. For the purpose of this paper, a total of 61 functional and 42 behavioral competencies were identified; these 103 competencies were considered features. In addition, six organizational performance metrics were identified as outputs (i.e., safety performance, market share, liquidity, profitability, revenue growth, and company image/reputation). A pairwise comparison between organizational performance metrics was conducted using the analytic hierarchy process (AHP) to determine their relative weights in order to obtain a single performance metric representative of organizational performance. The relative weights of the performance metrics are computed using the AHP as follows:

1. Develop the pairwise comparison matrix: Compare the metrics in pairs by giving them numbers on a scale based on Saaty's (1980) fundamental scale [25].

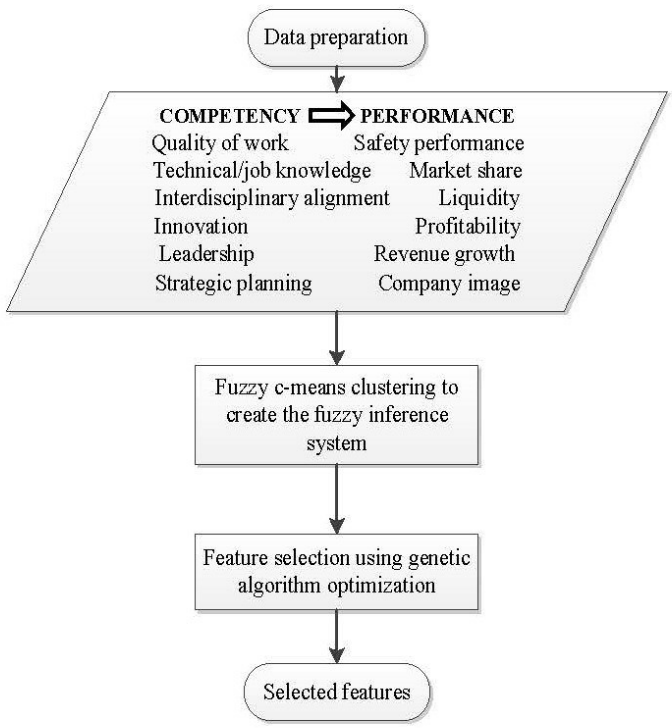

Fig. 1 Feature selection procedures for organizational competencies impacting performance

2. Determine the relative weights: First, normalize the pairwise comparison matrix by adding the values in each column of the pairwise comparison matrix and then dividing each cell by the total of the column. Then, determine the relative weights by calculating the average value of each row of the normalized matrix.

3. Check the consistency of the pairwise comparison matrix: The consistency ratio (*CR)* is a measure to control the consistency of the pairwise matrix. It is given as: $CR = CI/RI$, where *CI* is the consistency index and *RI* is the random consistency index. The *CI* is calculated as $CI = (\lambda_{max}–n)/(n–1)$, where λ_{max} is the maximum eigenvalue and n is the dimension of the pairwise comparison matrix. The *RI* measures how consistent the judgments have been relative to large samples of purely randomly generated comparison matrices [30]. The *RI* is dependent on the size of the pairwise comparison matrix and can be found from the *RI* tables [25]–[27].

For the pairwise comparison of the six organizational performance metrics (i.e., n = 6), the *RI* value obtained from the *RI* tables is 1.24. The maximum eigenvalue calculated is $\lambda_{max} = 6.4575$, where $CI = 0.092$. The *CR* of the pairwise comparison matrix for the AHP was checked and found to be 0.074, which is less than 0.1 and therefore consistent. Profitability and safety performance are the metrics with the highest relative weights, at 0.322 and 0.310, respectively. The relative weights of the remaining metrics i.e., liquidity, market share, reputation/company image, and revenue growth are calculated to be 0.111, 0.097, 0.082, and 0.078, respectively. Then, the performance metrics were aggregated using a weighted average to obtain a single performance metric. However, the major challenge of construction engineering and management research is the presence of constraints on getting adequate data (i.e., a limited number of data instances). Accordingly, a total of 56 data instances were randomly generated using integer values between 1 and 7 for the input features (i.e., the impact of competencies on performance) and integer values between 1 and 5 for the outputs (i.e., the measure of performance metrics).

C. Fuzzy c-means clustering to develop the fuzzy inference system

FCM is a method of fuzzy clustering in which each data point belongs to two or more clusters to a certain degree (i.e., partial belonging) [28], [29]. The generated data set was normalized and used for FCM clustering in order to develop a Mamdani-type FIS. Among the 56 data instances generated during data preparation, 46 were used for training while the remaining were used for testing. The FIS was developed with the training data, using the 103 competencies as input and the weighted average performance metric determined using the AHP as output. The accuracy of the FIS was checked using the root mean square error (RMSE) between predicted output values and the actual output values of the testing data. A combination of different values for the number of clusters and the fuzzification coefficient for the FCM clustering were used to develop the FIS. After conducting multiple iterations based on the RMSE, the optimum FCM parameters with minimum RMSE were found to be five clusters and a fuzzification coefficient of 2.75. Then, FS was conducted using GA optimization on the FIS.

D. Feature Selection using a genetic algorithm

The FS was performed using GA optimization, which in turn employed the FIS that was developed using optimum FCM parameters in the previous section. Population-based algorithm approaches widely used in FS are based on the GA [11], [20]. The GA is a stochastic global search optimization that attempts to arrive at optimal solutions through a process similar to biological evolution [28]–[30]. The GA operates on a population of potential solutions by applying the principle of survival of the fittest in addition to crossbreeding and mutation to generate better solutions from a pool of existing solutions [30], [31]. The GA evolutionary cycle starts with a randomly selected initial population (i.e., organizational competencies) and evaluates them using the fitness function. Changes to the population occur through the processes of selection based on fitness and alteration using crossover and mutation [30], [31].

IV. Results and Discussion

The MATLAB programming language was used to develop a code for finding the optimum value for clusters c and fuzzification coefficient m, performing FCM on input-output data. Accordingly, the optimum values determined for c and m are five and 2.75 respectively. Applying the optimum FCM parameters, an FIS was developed using the *genfis* options of MATLAB. Then FS was conducted using GA optimization on the FIS. The fitness function for the GA optimization is the RMSE. The GA selects features (i.e., competencies used as input) by optimizing the RMSE between the output of the FIS (i.e., predicted output) and the output from the actual (testing) data as a fitness value. The crossover and mutation probabilities were set as 0.8 and 0.1, respectively, while the number of generations was 100. After performing the GA optimization, a total of 16 competencies were selected out of 103 competencies with the best and mean fitness value (RMSE) of 0.035 and 0.037, respectively.

To validate the results, GA optimization for selecting the features was repeated using three and seven clusters, respectively, for the same fuzzification coefficient. As shown in Table II, a total of 13 and 22 competencies were selected for $c = 3$ and $c = 7$, respectively. GA optimization using the FIS with three clusters returned minimum features, but its predictive performance was poor (i.e., higher RMSE) compared to $c = 5$. It is also evident from Table II that the mean fitness value (RMSE) for $c = 7$ was marginally better than the value when $c = 5$. FS performed for the experimental case study using GA optimization makes it possible to reduce the dimensionality of the input features from 103 to 16 competencies, such as staff development and training, coordination and cooperation, project quality management, project cost management, resource management, effectiveness, attention to detail, and professionalism. Therefore, the 16 competencies selected as a representative subset of the original competencies can be used to develop a concise predictive model with good interpretability and generalization.

TABLE II. Results of FS using GA for different numbers of FCM clusters

No. of clusters	**No. of selected competencies (features)**	**Fitness value (RMSE)**	
		Best value	***Mean value***
3	13	0.0037	0.0039

No. of clusters	No. of selected competencies (features)	Fitness value (RMSE)	
		Best value	***Mean value***
5	16	0.0034	0.0036
7	22	0.0034	0.0035

The experiment is repeated following the approaches of [27] and [28]; that is, FS using GA with k-nearest-neighbor-based (KNN-based) classification (K-means clustering) for comparison. The 103 competences are treated as an input feature while the output (the weighted average performance metric) is considered class information for K-means clustering. First, a model is created by setting the values for KNN as 3, 5, and 7, respectively. Then, FS is performed using GA optimization on the model that was created using KNN. The result is shown in Table III, where the optimum features of 16 competencies are selected when KNN = 5. The fitness value (error) for the KNN model is marginally higher than the FIS model developed using FCM. Therefore, the 13 competencies selected for the FIS give a better result with optimum features that help to develop a compact and concise model with better interpretation and generalizability.

TABLE III. RESULTS OF FS USING GA FOR DIFFERENT NUMBERS OF KNN CLUSTERS BASED ON [30]

No. of clusters	No. of selected competencies (features)	Fitness value (RMSE)	
		Best value	***Mean value***
3	27	0.0037	0.0046
5	16	0.0054	0.0062
7	19	0.0054	0.0061

V. CONCLUSIONS

FS helps with the identification of good feature subsets related to a target concept (i.e., organizational competencies) characterized by highly dimensional features and limited data instances. This paper presented a review of FS, including the commonly used filter, wrapper, and embedded methods. A comparison of FS methods on the basis of their accuracy and computational performance was made.

An experimental case study was developed for FS that had 103 organizational competencies as input feature space and six organizational performance indicator metrics as an output. A total of 56 data instances were generated randomly to replicate the problem characterized by highly dimensional data with limited data instances. The six organizational performance indicator metrics were aggregated using the AHP to obtain a single performance metric that was used to develop the FIS for GA optimization. Then, FS was conducted using GA optimization on the FIS to select representative features among the original 103 competencies. The GA optimizes the RMSE between the predicted output of the FIS and the actual (test) data. As a result, a total of 16 competencies were selected using the optimized values of FCM. The experiment was repeated using the KNN classification method. The comparison of the result showed that GA optimization on FIS developed using FCM gives a better result compared to GA optimization on KNN classification. The case study showed that FS using GA optimization is effective for reducing the dimensionality of features from 103 competencies and providing a representative features (i.e., 16 competencies). It is evident that developing a predictive model using selected competencies rather than the original set of competencies significantly reduces the model's complexity and computational speed as well as enabling better interpretation and generalization. This paper contributes to the body of research by presenting an approach to FS that uses GA optimization on an FIS developed using FCM clustering and which is applicable to construction engineering and management problems characterized by highly dimensional features and few data instances.

ACKNOWLEDGMENT

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (NSERC RGPIN-2018-05174), which is held by Dr. Aminah Robinson Fayek. The authors gratefully acknowledge the financial support provided by NSERC.

REFERENCES

[1] A.P.C. Chan and A.P.L. Chan, "Key performance indicators for measuring construction success," Benchmarking: An International Journal, vol. 11, no. 2, pp. 203–221, 2004, doi: 10.1108/14635770410532624.

[2] A.S. Hanna, M.W. Ibrahim, W. Lotfallah, K.A. Iskandar, and J.S. Russell, "Modeling project manager competency: An integrated mathematical approach," J. Constr. Eng. Manage., vol. 142, no. 8, 04016029, 2016, doi: 10.1061/CO.1943-7862.0001141.

[3] M. Radujković, M. Vukomanović, and I.B. Dunović, “Application of key performance indicators in south-eastern European construction,” J. Civ. Eng. Manage., vol. 16, no. 4, pp. 521–530, 2010, doi: 10.3846/jcem.2010.58.

[4] A. Chakravarty, R. Grewal, and V. Sambamurthy, “Information technology competencies, organizational agility, and firm performance: Enabling and facilitating roles,” Information Systems Research, vol. 2, no. 4, pp. 976–997, 2013, doi: 10.1287/isre.2013.0500.

[5] S. Loufrani-Fedida and S. Missonier, “The project manager cannot be a hero anymore! Understanding critical competencies in project-based organizations from a multilevel approach,” Interna. J. Proj. Manage., vol. 33, no. 6, pp. 1220–1235, 2015, doi: 10.1016/j.ijproman.2015.02.010.

[6] M.N. Omar and A.R. Fayek, “Modeling and evaluating construction project competencies and their relationship to project performance,” Automation in Construction, vol. 69, pp. 115–130, 2016, doi: 10.1016/j.autcon.2016.05.021.

[7] B. Giel and R.R.A. Issa, “Framework for evaluating the BIM competencies of facility owners,” J. Manage. Eng., vol. 32, no. 1, 2016, Art. no. 04015024, doi: 10.1061/(ASCE)ME.1943-5479.0000378.

[8] G.G. Tiruneh and A.R. Fayek, “A framework for modeling organizational competencies and performance,” ASCE Constr. Res. Congr., pp. 712–722, 2018, doi: 10.1061/9780784481271.069.

[9] G.G. Tiruneh and A.R. Fayek, “Identifying construction organizational competency measures and performance indicator metrics,” 6th CSCE Construction Specialty Conference, Vancouver, BC, Canada, May 31–June 3, 2017.

[10] P. Mitra, C.A. Murthy, and S. K. Pal, “Unsupervised feature selection using feature similarity,” IEEE Trans. Pattern Analysis Machine Intel., vol. 24, no. 3, pp. 301–312, 2002.

[11] S.S.S. Ahmad and W. Pedrycz, “Data and feature reduction in fuzzy modeling through particle swarm optimization,” Applied Computational Intelligence and Soft Computing, vol. 2012, 2012, doi: 10.1155/2012/347157.

[12] V. Kumar and S. Minz, “Feature selection: A literature review,” Smart Computing Review, vol. 4, no. 3, pp. 211–229, 2014, doi: 10.6029/smartcr.2014.03.007.

[13] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, “Recent advances and emerging challenges of feature selection in the context of big data,” Knowledge-Based Systems, vol. 86, pp. 33–45, 2015, doi: 10.1016/j.knosys.2015.05.014.

[14] R. Kavitha and E. Kannan, “An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining,” IEEE Int. Conf. Emerg. Trends Eng., Tech. Sci., 2016, doi: 10.1109/ICETETS.2016.7603000.

[15] K. Liu and N. El-Gohary, “Feature discretization and selection methods for supporting bridge deterioration prediction,” ASCE Constr. Res. Congr., 2018, pp. 413–423, 2018, doi: 10.1061/9780784481295.042.

[16] G. Taskin, H. Kaya, and B. Lorenzo, “Feature selection based on high dimensional model representation for hyperspectral images,” IEEE Trans. Image Proc., vol. 26, no. 6, pp. 2918–2928, 2017, doi: 10.1109/TIP.2017.2687128.

[17] G. Taskin, H. Kaya, and B. Lorenzo, “A novel method for feature selection with random sampling HDMR and its application to hyperspectral image classification,” IEEE Int. Geosci. Remote Sensing Symp., 2015, pp. 1108–1111.

[18] A.H. Shahana and V. Preeja, “Survey on feature subset selection for high dimensional data,” IEEE Int. Conf. Circuit, Power Comp. Tech., 2016, doi: 10.1109/ICCPCT.2016.7530147.

[19] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[20] G. Chandrashekar, and F. Sahin, “A survey on feature selection methods,” Computers and Electrical Engineering, vol. 40, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[21] C. Liu, W. Wang, Q. Zhao, X. Shen, and M., Konan “A new feature selection method based on a validity index of feature subset,” Pattern Recognition Letters, vol. 92, pp. 1–8, 2017, doi: 10.1016/j.patrec.2017.03.018.

[22] A.A. Tsehayae and A.R. Fayek, “Developing and optimizing context-specific fuzzy inference system-based construction labor productivity models,” J. Constr. Eng. Manage., vol. 142, no. 7, 2016, Art. no. 04016017, doi: 10.1061/CO.1943-7862.0001127.

[23] S. Saitta, P. Kripakaran, B. Raphael, and I.F.C. Smith, “Feature selection using stochastic search: An application to system identification,” J. Comput. Civ. Eng., vol. 24, no. 1, pp. 3–10, 2010, doi: 10.1061/CP.1943-5487.0000003.

[24] G.G. Tiruneh and A.R. Fayek, “Focus group study to evaluate and verify construction organizational competency and performance measures,” J. Constr. Eng. Manage., unpublished.

[25] T. L. Saaty, “Decision making with the analytic hierarchy process,” Int. J. Services Sciences, vol. 1, no. 1, pp. 83–98, 2008.

[26] E. Mu and M. Pereyra-Rojas, “Understanding the analytic hierarchy process,” in Practical Decision Making. SpringerBriefs in Operations Research, 2017, pp. 7–22, doi: 10.1007/978-3-319-33861-3_2.

[27] T. L. Saaty, Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World, Third Revised Edition. Pittsburgh: RWS Publications, 2012.

[28] M. Alata, M. Molhim, and A. Ramini, “Using GA for optimization of the fuzzy c-means clustering algorithm,” Res. J. Appl. Sci. Eng. Technol., vol. 5, no. 3, pp. 695–701, 2013.

[29] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems. CRC Press, Taylor & Francis Group, 2013.

[30] B. Oluleye, A. Leisa, J. Leng, and D. Dean, “Zernike moments and genetic algorithm: Tutorial and application,” British Journal of Mathematics and Computer Science, vol. 4, no. 15, pp. 2217–2236, 2014.

[31] T.P. Pushpavathi, V. Suma, and V. Ramaswamy, “Defect prediction in software projects-using genetic algorithm based fuzzy c-means clustering and random forest classifier,” International Journal of Scientific & Engineering Research, vol. 5, no. 9, pp. 888–898, 2014.