Framework for Integrating an Artificial Neural Network and a Genetic Algorithm to Develop a Predictive Model for Construction Labor Productivity

Sara EBRAHIMI[1], Mohammad RAOUFI[2], and Aminah Robinson FAYEK[3]

[1] MSc Student and Graduate Research Assistant, Hole School of Construction Engineering, Department of Civil and Environmental Engineering, University of Alberta, 7-203 Donadeo Innovation Centre for Engineering, 9211 116 Street NW, Edmonton, Alberta, T6G 1H9, Canada, email: eb4@ualberta.ca

[2] Postdoctoral Fellow, Hole School of Construction Engineering, Department of Civil and Environmental Engineering, University of Alberta, 7-381 Donadeo Innovation Centre for Engineering, 9211 116 Street NW, Edmonton, Alberta, T6G 1H9, Canada, email: mraoufi@ualberta.ca

[3] Professor, Director of the Construction Innovation Centre, Tier 1 Canada Research Chair in Fuzzy Hybrid Decision Support Systems for Construction, NSERC Industrial Research Chair in Strategic Construction Modeling and Delivery, Ledcor Professor in Construction Engineering, Hole School of Construction Engineering, Department of Civil and Environmental Engineering, University of Alberta, 7-232 Donadeo Innovation Centre for Engineering, 9211 116 Street NW, Edmonton, Alberta, T6G 1H9, Canada, PH: (780) 492-1205, email: aminah.robinson@ualberta.ca

ABSTRACT

Construction labor productivity (CLP) is one of the most important factors in the construction industry, as it has a direct effect on a company's efficiency and profitability. The accurate prediction of CLP is essential for effective decision-making prior to project execution, and continuous tracking and improvement of productivity over a project life cycle is necessary for its success. The objective of this paper is to develop a framework to help construction organizations predict and measure construction productivity, leading to improved project performance in terms of cost, time, and quality. CLP is affected by numerous factors, including the high-dimensional factors that result from a large number of model input variables and which often impose a high computational cost and the risk of overfitting of data. Therefore, it is necessary to use feature selection methods to reduce the dimensionality of CLP data. This paper proposes a framework that integrates an artificial neural network (ANN) and a genetic algorithm (GA) for feature selection. The proposed framework is used to develop a predictive model for CLP using features selected because they provide the best prediction of CLP. The ability of GAs to generate an optimal feature subset in combination with the superior accuracy of ANNs is a unique advancement that this framework offers for improving the prediction of labor productivity. The developed model can predict productivity and specify which factors are most predictive of CLP. The contributions of this paper are (1) the development of a framework that uses an integrated ANN and GA as a wrapper method for selecting the features with the most influence on CLP and (2) the development of an improved predictive model that can be used to both predict and measure CLP.

INTRODUCTION

Since many activities in the construction industry are labor dependent, improving labor productivity is key for improving project performance. Construction labor productivity (CLP) significantly impacts a company's profitability and project cost, and construction organizations therefore need a predictive model of activity-level CLP that helps them understand which factors affect labor productivity (Moselhi and Khan 2012). Numerous factors, both subjective (e.g., foreman skill and task complexity) and objective (e.g., crew size), have been identified that affect CLP, causing complex variability (Tsehayae and Fayek 2014; Raoufi and Fayek 2018; Hamza et al. 2019). Therefore, providing a predictive model for CLP requires complex mapping of the affecting factors (Heravi and Eslamdoost 2015). A large number of inputs and high-dimensional data may present different problems, such as reduced accuracy and increased complexity (Piao and Ryu 2017). To overcome these problems and find the factors with the most influence on CLP, feature selection methods are used. In data mining, feature selection is a necessary preprocessing approach for identifying a relevant subset for classification. The aim of feature selection is to quickly develop prediction models with better performance (Piao and Ryu 2017).

Past research that used both filter methods and wrapper methods indicates that the wrapper method produces better results in feature selection (Alolfe et al. 2009). However, current CLP studies are limited in their use of feature selection methods, as they use only filter methods.

In this paper, a framework for feature selection that uses an ANN and a GA as a wrapper method is developed and used to produce a predictive model for CLP. The framework integrates the ANN and the GA for feature selection in order to find the optimal feature subset by minimizing the fitness error of the ANN. Then, by employing the selected features in the ANN as inputs with CLP as the output, a model for predicting CLP is developed.

This approach predicts and measures CLP using the most influential factors, which are selected by employing the neural-genetic algorithm as a combination of the ANN and GA for selecting the best predictive CLP factors. Using both the abovementioned feature selection algorithm and a database for CLP factors enables the identification of the factors with the most influence on CLP so they can be taken into account on various construction projects.

This paper is organized as follows: First, a review of past research on feature selection methods, past methods used to select CLP factors, and the integration of an ANN and a GA is presented. Then, a neural-genetic algorithm for selecting the best predictive factors of CLP is described. Using the selected features, a predictive model for labor productivity is developed. Next, a case study and the results of implementing the algorithm are presented along with the predictive model. Finally, conclusions and future research are presented.

LITERATURE REVIEW

Feature selection is the process of identifying and removing irrelevant and redundant data (Hall 1999). It focuses on choosing a subset of the input features that efficiently represents the input data while decreasing irrelevant features or noise effects and providing a relatively accurate prediction of results. The main benefits of feature selection are that (1) it decreases the amount of data needed to achieve learning, (2) it enhances the predictive accuracy of models, (3) it reduces model execution time because there are fewer inputs, and (4) it allows learned knowledge to be easily understood because it is more compact (Hall 1999). Filter and wrapper methods are the main approaches for feature selection (Yao et al. 2015). Filter methods are independent of learning algorithms and choose best features based on some of the statistical properties of data, such as their correlation coefficients. Because a small number of features are used for classification in filter methods, the computation cost is low, which is the main advantage of these methods. However, a small number of features, even if they are the “best” ones, does not guarantee high classification accuracy (Cover 1974). Furthermore, most filter methods are only suitable for developing mathematical equations by the statistical regression method (Guyon et al. 2008; Gerami Seresht and Fayek 2018; Raoufi and Fayek 2018a). Wrapper methods, on the other hand, use the accuracy of a learning algorithm as a criterion for selecting useful features (Yao et al. 2015), and they explore the feature space to score feature subsets according to their predictive power. Wrapper methods are therefore a more effective means of constructing a predictive model than filter methods because they are tuned to the specific interaction between a learning algorithm and its training data (Ahmad et al. 2015; Aličković and Subasi 2017). However, their application is limited because of the high computational complexity that occurs when numerous feature sets are considered (Piao and Ryu. 2017).

In the construction discipline, there are some studies that use filter feature selection for reducing the number of factors influencing CLP. Tsehayae and Fayek (2014) identified a total of 169 parameters influencing CLP in building and industrial projects. As the 169 input parameters and seven process variables result in a high-dimension feature space, Tsehayae and Fayek (2016) used a correlation-based feature selection (CFS) algorithm, which is a filter method, and proposed a predictive model that uses a fuzzy inference system. Although the CFS algorithm is suitable in that it has the ability to deal with a high dimension of input space and a small number of data instances, using a wrapper method is more appropriate for predictive modeling using artificial intelligence (AI) techniques, such as fuzzy inference systems and ANNs, because of its superior performance (Piao and Ryu 2017). Therefore, this research proposes the integration of an ANN and a GA as a wrapper method for CLP feature selection.

The integration of the ANN with the GA has not been investigated in CLP studies in the construction domain. However, in other domains, such as medical research, the integration of ANNs with GAs for feature selection has been studied. For example, the technique of combining ANN parameters that are simultaneously optimized by the GA was proposed by Verma and Zhang (2007) and Ahmad et al. (2015) to implement feature selection for diagnosing breast cancer. The classification rates achieved in both studies were promising and showed better results than most previous studies that used filter methods.

GAs, inspired by the natural selection process, are strong evolutionary optimization algorithms that search for the best subset of system parameters for the development of an accurate predictive model. GAs have been applied successfully for feature selection by various researchers (Gerami Seresht and Fayek 2018). The integration of a GA and an ANN (e.g., in Verma and Zhang 2007) has recently attracted wide attention. The objective of this paper is to present a neural-genetic algorithm for finding the most influential features and developing a predictive model for CLP.

METHODOLOGY TO DEVELOP A NEURAL-GENETIC ALGORITHM FOR FEATURE SELECTION

This paper presents a neural-genetic algorithm for finding the most influential factors from a number of existing features that affect CLP in order to develop a model for predicting CLP.

Figure 1 shows the conceptual integration of a GA and an ANN, where each individual in the population indicates a candidate solution for selecting the feature subset. If there are n features affecting CLP, there are 2^n possible feature subsets. Each feature subset is called a "chromosome" and contains n genes, which can have one of two values. A value of 1 indicates that the corresponding feature has been chosen for predicting CLP, and a value of 0 means that the feature has not been selected.

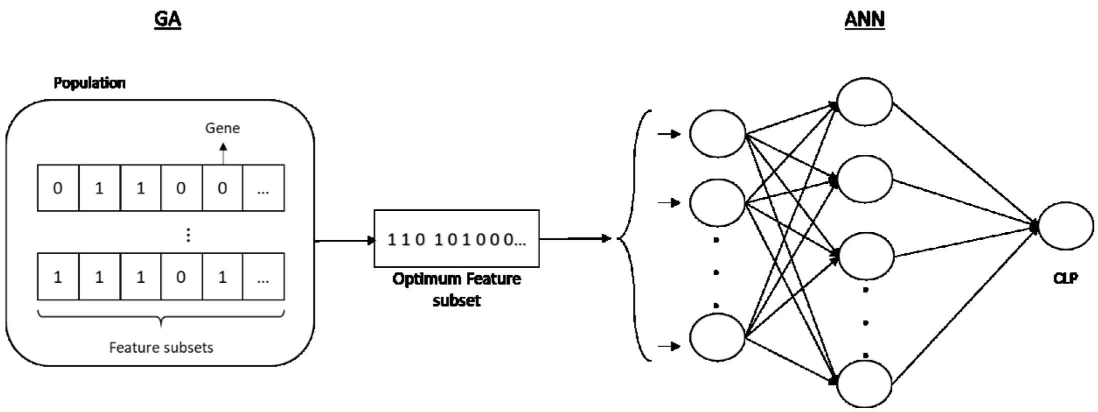

Figure 1. Conceptual integration of GA and ANN.

An overview of the proposed framework is shown in Figure 2, which presents the process of integrating the GA and the ANN for feature selection. There five steps to performing this integration, described in detail in the following paragraphs.

As shown in Figure 2, in the first step, the algorithm generates the random initial population of chromosomes. Each individual in the population represents an available solution to the feature subset selection problem.

In the second step, the selected features are the inputs of the ANN. In the neural network, the number of hidden layer nodes is calculated using Equation (1), as the appropriate hidden layer size is calculated based on the number of inputs and outputs (Heaton 2008).

$$Number\ of\ nodes\ =\ 2 * \sqrt{inputs + outputs} \qquad (1)$$

In the third step, the training set, validation set, and testing set are built from the CLP dataset collected by Tsehayae and Fayek (2014). Then, using the training set, the process of training the network is started. To avoid overtraining in ANN, the error of the validation set is considered during this process. If the error grows for five iterations consecutively, then the training stops.

The fourth step is the fitness calculation process, wherein the validation set is used to simulate the network and calculate the error by using the root mean square error (RMSE) based on Equation (2).

$$Error = \sqrt{\frac{1}{n}\sum_{i=1}^{p}(A_i - T_i)^2} \quad (2)$$

where p, A_i and T_i are the number of output nodes, the actual output value of the ith output node, and the target output value of the ith output node, respectively. In this paper, there is one output node, which is CLP. A better fitness of the ANN requires a smaller error.

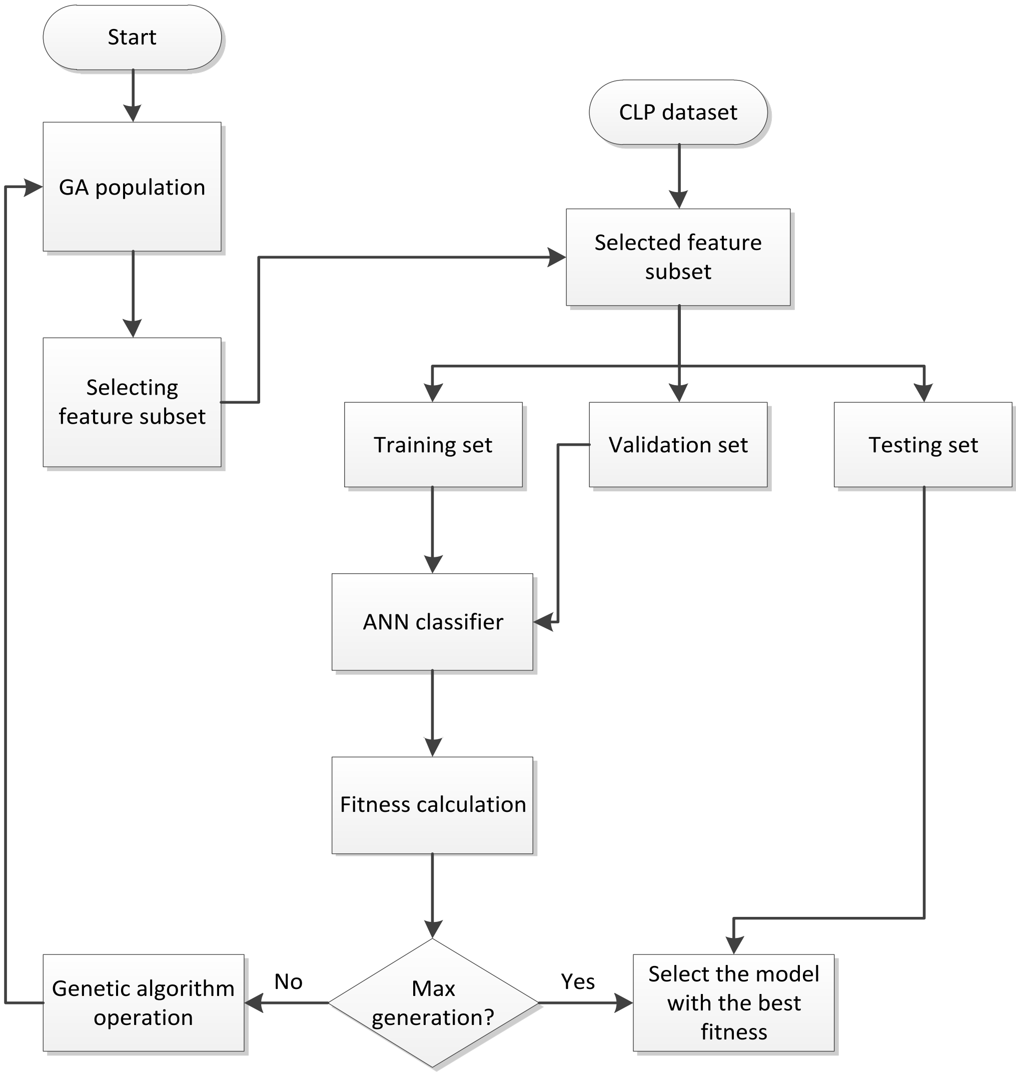

Figure 2. Overview of the CLP feature selection algorithm.

The fifth step is a GA operation, which consists of the following process:

(a) Selection: A roulette wheel selection strategy is used to choose the individual probabilistically to form a parent whose number is equal to the population size minus the elitism number. If f_i is the fitness of individual i in the population, the probability of being selected is given by Equation (3), where N is the population size.

$$P_i = i\frac{f_i}{\sum_{j=1}^{N} f_j} \quad (3)$$

(b) Crossover and mutation: Crossover is the process used to specify which two chromosomes will create a new offspring chromosome (Bean 1994). The mutation operation changes one or more genes in a chromosome from its initial state. All the chromosomes after the crossover operation will go through a mutation operation and a new offspring is produced. In this research, single point binary crossover and binary mutation are performed.

(c) Elitism: Another process in the GA algorithm is elitism. Elitism involves copying a small proportion of the fittest candidates, unchanged, into the next generation. In this paper, the three best chromosomes are selected to be part of the population in the next generation.

(d) Fitness function: The GA optimization method minimizes the value of a fitness function, which is shown as Z and calculated for each chromosome. The Z function is defined by Equation (4), where b is a coefficient of the number of selected features (nf). In this study, we consider "b" to be equal to 0.008 in order to have fewer selected features.

$$Z = Error * (1 + b * nf) \quad (4)$$

After these processes, once the final generation with the best fitness value is reached, the iteration stops, and the feature subset that is chosen as a final solution is the one that is the best predictor of CLP among all feature subsets. Accordingly, a predictive model for CLP can be developed, which has the minimum fitness error.

CASE STUDY

To illustrate the proposed method of feature selection and construct the predictive model for CLP, a case study was conducted. Tsehayae and Fayek (2014) identified a total of 96 activity-level sub-parameters. In this case study, all 20 activity-level factors, identified by Tsehayae and Fayek (2014), that showed non-zero variance in data were considered inputs to the feature selection algorithm (Table 1), and CLP was the output of the predictive model. A total of 92 data instances were used. The aim of feature selection in this case study is to identify the most influential features among the 20 activity-level factors to be able to quickly develop a predictive model of CLP.

The proposed algorithm was developed in the MATLAB 9.6 environment. The backpropagation technique, which is one type of ANN learning algorithm, was used due to its fast execution and simple implementation in MATLAB. The output and the hidden nodes' activation functions were pure linear and hyperbolic tangent, respectively. The output layer consisted of one output node, which is CLP. Table 2 shows the GA parameter settings, which were based on Zhuo et al. (2008). In order to make data consistent across all tables, feature normalization was required. This approach uses normalized data, which are real numbers in the range 0–1. In this paper, 70% of the CLP dataset was used for training, the next 15% was used for validation, and the last 15% was used for testing. The validation set was used to calculate the overall fitness of the network and choose the best network, and the testing set was used to achieve the desired test accuracy of the selected neural network. The early stopping method, which defines the maximum number of iterations before overfitting begins, was used to avoid overfitting of the network (Ebrahimi Kahou et al. 2015; Gal and Ghahramani 2016).

Table 1. CLP factors.

No	Factor
f1	Crew size
f2	Craftsperson education
f3	Craftsperson technical training
f4	Crew composition
f5	Crew experience (seniority)
f6	Number of languages spoken
f7	Cooperation among craftspersons
f8	Treatment of craftspersons by foreman
f9	Craftsperson motivation
f10	Craftsperson fatigue
f11	Team spirit of crew
f12	Fairness of work assignments
f13	Crew participation in foreman decision-making process
f14	Crew flexibility
f15	Availability of task materials
f16	Quality of task materials
f17	Sharing of tools
f18	Working condition (noise)
f19	Location of work scope (distance)
f20	Fairness in performance review of crew by foreman

Table 2. GA parameter settings.

Parameter	
Population size	50
Crossover probability	0.8
Elitism size	3
Mutation probability	0.2
Maximum generation	40

RESULTS

As a result of implementing the neural-genetic algorithm, a total of seven features were selected as the factors with the most influence on CLP, and they were used to develop a predictive model. These seven features were crew size (f1), craftsperson technical training (f3), cooperation among craftspersons (f7), craftsperson motivation (f9), craftsperson fatigue (f10), fairness of work assignments (f12) and sharing of tools (f17). Table 3 shows the results of implementing the neural-genetic algorithm. As shown in Table 3, the final error of fitness calculation of the ANN was 0.0107, which was calculated using Equation (2). Based on Equation (4), Z represents the minimum amount of fitness function. As the number of selected features (i.e., inputs) was seven and the number of outputs was one, the number of hidden layer nodes in the selected model was calculated to be six using Equation (1).

Table 3. Results of conducting neural-genetic algorithm.

	Results
Selected features	f1, f3, f7, f9, f10, f12, f17
Error	0.0107
Z	0.0111
Hidden layer size	6

These selected features were the inputs of the achieved predictive model with the best fitness, which is shown in Figure 3. The limitation of the wrapper method in terms of computational complexity in dealing with numerous feature sets does not occur in this case because the utilized CLP dataset consists of 20 features.

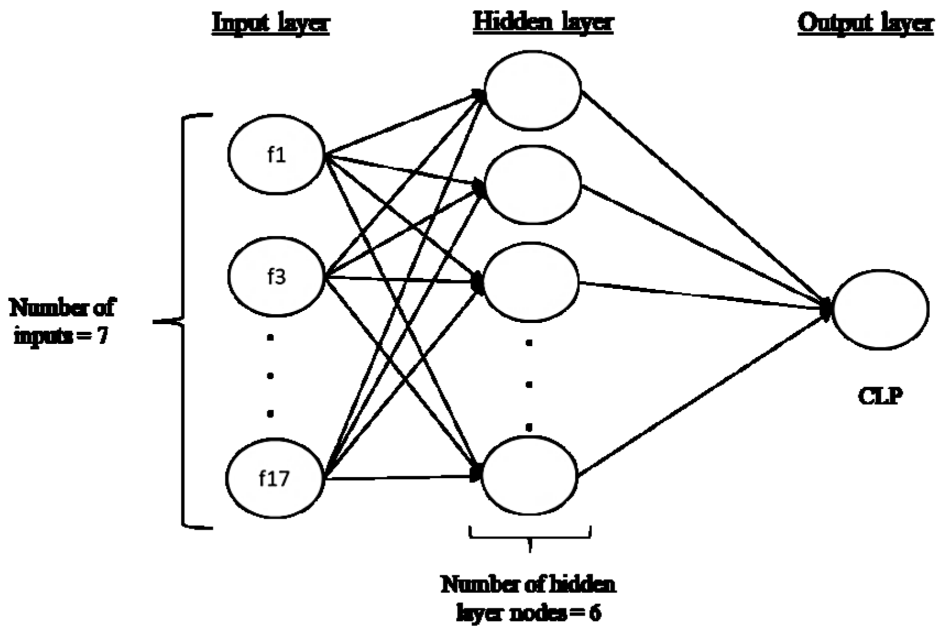

Figure 3. The predictive model for CLP.

One of the selected features was cooperation among craftspersons (f7), which is supported by Tsehayae and Fayek (2014), whose work identified "good cooperation between craftsmen in a crew" as a top parameter having a positive effect on CLP. In addition, Jergeas (2009) found "labor relations" to be a target for CLP improvement. Tsehayae (2015) identified 27 identical input variables for CLP in a total four contexts (i.e., industrial, warehouse, high-rise, and institutional building) by using a CFS algorithm including crew size, craftsperson on-job training, craftsperson motivation, craftsperson fatigue, and fairness of work assignments, which are found in our results (f1, f3, f9, f10 and f12). Therefore, the selected features in this paper are in agreement with the results of past studies on productivity feature selection. However, the comparison of the results of this study with past literature indicates that the proposed framework can better identify predictive features of CLP. Tsehayae and Fayek (2016) obtained 2.515% as the RMSE value, while in this paper RMSE value is 1.070%. Future research will focus on collecting a larger data set from different organizations and project contexts to expand the scope of applicability of the developed algorithm.

CONCLUSIONS AND FUTURE RESEARCH

The main aim of this paper was to develop a predictive model for measuring and predicting CLP using a GA algorithm and an ANN. After performing a literature review to investigate past research on feature selection methods and the importance of CLP, a methodology for constructing a predictive model for CLP was developed. This paper illustrated the neural-genetic framework for both feature selection and presenting a predictive model. By implementing the developed framework on a real case, the features that were the best predictive factors of CLP were identified, and the results were compared to past research and found to be consistent with previous results. The achieved error in this paper indicates an improvement of the predictive model in comparison to past studies. The contributions of this paper are (1) the development of a framework that uses an ANN and a GA as a wrapper method for feature selection to select the parameters with the most influence on CLP and (2) the development of an improved model for predicting and measuring CLP. The results of this work will improve the prediction of CLP. Better identification of the factors with the most influence on CLP can lead to more effective management of CLP and project performance. The findings of this paper also provide a basis for future research work, including modeling CLP based on feature selection on all identified influencing factors with actual data. Modeling multifactor construction productivity, which includes labor, material, and equipment, can be done in future works by using the proposed framework. Future research can also focus on implementing other AI techniques, such as the combination of a GA and a neuro-fuzzy system, to focus on subjective factors affecting CLP and comparing the accuracy of predictive models with different AI techniques.

REFERENCES

Ahmad, F., Isa, N. A. M., Hussain, Z., Osman, M. K., and Sulaiman, S. N. (2015). “A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer.” *Pattern Analysis and Applications*, 18(4), 861-870.

Aličković, E., and Subasi, A. (2017). “Breast cancer diagnosis using GA feature selection and Rotation Forest.” *Neural Computing and Applications*, 28(4), 753-763.

Alolfe, M. A., Mohamed, W. A., Youssef, A. M., Kadah, Y. M., and Mohamed, A. S. (2009). “Feature selection in computer aided diagnostic system for microcalcification detection in digital mammograms.” *Proc., 2009 National Radio Science Conference*, IEEE, 1-9.

Bean, J. C. (1994). “Genetic algorithms and random keys for sequencing and optimization.” *ORSA Journal on Computing*, 6(2), 154-160.

Cover, T. M. (1974). “The best two independent measurements are not the two best.” *IEEE Trans. Syst. Man Cybern*., SMC-4(1), 116-117.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. (2015). “Recurrent neural networks for emotion recognition in video.” In Proc., *the 2015 ACM on International Conference on Multimodal Interaction*, ACM, Seattle, Washington, USA, 467-474.

Gal, Y., and Ghahramani, Z. (2016). “A theoretically grounded application of dropout in recurrent neural networks.” In *Proc., 30th Annual Conference on Neural Information Processing Systems 2016 (NIPS 2016)*, Barcelona, Spain, 1019-1027.

Gerami Seresht, N., and Fayek, A. R. (2018). "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *J. Constr. Eng. Manage.*, 144(9), 04018091.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). "Feature extraction: Foundations and applications." *Springer*.

Hall, M. A. (1999). "Correlation-based feature selection for machine learning."

Hamza, M., Shahid, S., Bin Hainin, M. R., and Nashwan, M. S. (2019). "Construction labour productivity: Review of factors identified." *Int. J. Constr. Manage.,* 1-13.

Heaton, J. (2008). "*Introduction to neural networks with Java.*" Heaton Research, Inc., St. Louis, USA.

Heravi, G., and Eslamdoost, E. (2015). "Applying artificial neural networks for measuring and predicting construction-labor productivity." *J. Constr. Eng. Manage.*, 141(10), 04015032.

Jergeas, G. (2009). "Improving construction productivity on Alberta oil and gas capital projects." A Report Submitted to: Alberta Finance and Enterprise.

Moselhi, O., and Khan, Z. (2012). "Significance ranking of parameters impacting construction labour productivity." *Construction Innovation,* 12(3), 272-296.

Piao, Y., and Ryu, K. H. (2017). "A hybrid feature selection method based on symmetrical uncertainty and support vector machine for high-dimensional data classification." *Proc., Asian Conference on Intelligent Information and Database Systems*, Springer, 721-727.

Raoufi, M., and Fayek, A. R. (2018a). "Key moderators of the relationship between construction crew motivation and performance." *J. Constr. Eng. Manage*., 144(6), 04018047.

Raoufi, M., and Fayek, A. R. (2018b). "Framework for identification of factors affecting construction crew motivation and performance." *J. Constr. Eng. Manage.,* 144(9), 04018080.

Tsehayae, A. A. (2015). "Developing and optimizing context-specific and universal construction labour productivity models."

Tsehayae, A. A., and Fayek, A. R. (2014). "Identification and comparative analysis of key parameters influencing construction labour productivity in building and industrial projects." *Can. J. Civ. Eng*., 41(10), 878-891.

Tsehayae, A. A., and Fayek, A. R. (2016). "Developing and optimizing context-specific fuzzy inference system-based construction labor productivity models." *J. Constr. Eng. Manage.*, 142(7), 04016017.

Verma, B., and Zhang, P. (2007). "A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms." *Applied Soft Computing*, 7(2), 612-625.

Yao, J., Mao, Q., Goodison, S., Mai, V., and Sun, Y. (2015). "Feature selection for unsupervised learning through local learning." *Pattern Recog. Lett.*, 53 100-107.

Zhuo, L., Zheng, J., Li, X., Wang, F., Ai, B., and Qian, J. (2008). "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine." *Proc., Geoinformatics 2008 and Joint Conference on GIS and Built Environment*, 71471J.