Developing Hybrid Artificial Intelligence Model for Construction Labour Productivity Prediction and Optimization

by

Sara Ebrahimi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Construction Engineering and Management

Department of Civil and Environmental Engineering

University of Alberta

Abstract

Construction labour productivity (CLP) is considered one of the most important parameters affecting the performance of construction projects. Therefore, modeling CLP is a crucial step in construction projects. Accurate prediction of CLP helps in effective planning, cost estimating, and productivity improvement before and during construction project execution. Numerous factors affect CLP and cause complexity in predicting and modeling labour productivity. Thus, CLP modeling and prediction are complex tasks, which can lead to high computational cost and overfitting of data. Since a large number of inputs and high-dimensional data may present different problems, such as reduced accuracy and increased complexity, it is necessary to reduce the dimensionality of CLP data and determine the factors that most influence CLP. This can be accomplished using dimensionality reduction methods, such as feature selection. Existing predictive models of CLP do not focus on dimensionality reduction methods appropriately, which causes reduced accuracy of CLP prediction.

This thesis presents a novel approach to predict and optimize CLP by applying hybrid feature selection (HFS), machine learning models, and particle swarm optimization (PSO) algorithm. HFS methods select the most predictive factors on CLP to reduce complexity and dimensionality of CLP. Selected factors are used as inputs to four machine learning models, namely adaptive neuro-fuzzy system (ANFIS), ANFIS-genetic algorithm (ANFIS-GA), random forest (RF), and artificial neural network (ANN) for CLP prediction. Results show that the RF model obtains better performance compared to the other three models. Finally, the integration of RF and PSO is developed to identify the maximum value of CLP and the optimum value of selected factors. The new hybrid model presented, named HFS-RF-PSO, is a CLP optimization-and-prediction approach that addresses the limitation of existing CLP prediction studies regarding the lack of

capacity to optimize CLP and its most influential factors in regard to a construction company's preferences, such as targeted CLP. Therefore, the main contributions of this thesis include (1) development of an HFS model to select the most predicting factors on CLP; (2) development and comparison of four different predictive models for CLP and identifying the most accurate model; and (3) development of the HFS-RF-PSO algorithm to identify the maximum value of CLP considering the minimum deviation from the targeted CLP value and also finding the optimum value of the selected.

The proposed HFS-RF-PSO model will help project managers predict, optimize, and improve the CLP value while taking into account the factors that are most predictive of CLP. The results of this thesis and implementation of the HFS-RF-PSO model will help project managers identify causes of low labour productivity, select and prioritize corrective measures to improve CLP. The model will also enable project managers to improve the reliability of predictions.

Preface

This thesis is an original work by Sara Ebrahimi. Chapter 4 and parts of Chapter 3 have been published as Ebrahimi, S., Fayek, A. R., and Sumati, V. (2021). “Hybrid Artificial Intelligence HFS-RF-PSO Model for Construction Labor Productivity Prediction and Optimization.” *Algorithms*, Multidisciplinary Digital Publishing Institute, 14 (7), page 214, published July 15, 2021. Moreover, parts of Chapter 3 have been submitted for publication as Ebrahimi, S.; Kazerooni, M., Sumati, V., and Fayek, A. R. (n.d.). “A predictive model for construction labour productivity using the integration of hybrid feature selection and PCA Methods,” in review, submitted to *Canadian Journal of Civil Engineering*. I was responsible for the major parts of the data collection, analysis, and composition of the manuscript. A. R. Fayek was the supervisory author and was involved with concept formation, composition, and editing of the manuscript.

Dedication

I dedicate this research to my beloved parents, Mahzad Akhavein and Abolfazl Ebrahimi, my great brother, Amirhossein Ebrahimi, and my amazing partner, Matin Kazerooni.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Aminah Robinson Fayek for her expertise, close follow-up, continuous intellectual support, encouragement, and the energy and time she dedicated to guiding me along the course of this research. I would have never been able to accomplish this thesis without her assistance and her support, especially during the Covid-19 pandemic situation.

I would like to acknowledge postdoctoral fellows Dr. Nima Gerami Seresht, Dr. Mohammad Raoufi, Dr. Phuong Nguyen, and Dr. Sumati Vuppuluri for their advice and assistance. I would like to thank my colleagues at the University of Alberta, Sahand Somi, Seyed Hamed Fateminia, Yisshak Gebretekle, and Nebiyu Siraj for all their help during my studies. I would like to thank my colleague, Renata Brunner Jass for her great works on editing my research papers.

Finally, I would like to express my deepest gratitude to my parents, Mahzad and Abolfazl, my brother, Amirhossein and my partner, Matin for their support, encouragement, and inspiration during my research.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

As the construction industry makes a significant contribution to global gross domestic product, sustaining construction productivity is essential to economic growth (El-Gohary et al. 2017). In 2015–2019 in Canada, the construction industry contributed to about 7.3% of the gross domestic product and also provided employment for about 7.6% of all employees (Statistics Canada 2019). The efficiency of construction systems is measured by using construction productivity. Poor productivity can cause cost overruns and schedule slippages on large, labour-intensive construction projects (Doloi 2008). In general, productivity can be defined as the ratio of inputs of the system (e.g., person-hours or cost) to its output (e.g., cubic meters of concrete placed). Based on Talhouni (1990), there are three types of measurements for construction productivity, namely single factor productivity (SFP), multifactor productivity (MFP), and total factor productivity (TFP). SFP measures productivity by using one resource input, while MFP uses any combination of labour, equipment, and materials as the resource inputs. TFP measures the construction productivity by using labour, materials, equipment, energy, and capital as five resource inputs. As predicting and measuring energy and capital inputs in activity or at the project level is difficult, measuring TFP can be inaccurate (Gerami Seresht and Fayek 2018; Loosemore 2014). Therefore, construction managers often use construction labour productivity (CLP), which is a SFP measure that utilizes labour as an only input (Eastman and Sacks 2008; Loosemore 2014; Tsehayae and Fayek 2016a). In this thesis, the focus is on CLP, which is defined as either the ratio of units of output to units of input or the ratio of units of input to units of output. In this study, CLP is defined as shown in Eq. (1.1), where higher values are better than lower values. For instance, the concrete activity installed quantity is measured in terms of total volume placed.

$$CLP = \frac{Output\ (Installed\ quantity)}{Total\ labor\ work{-}hours} \qquad (1.1)$$

Three different levels are considered for construction productivity. The first is economic-level productivity, which is suitable for industry-wide measurements of construction productivity. The second is project-level, which is focused on specific projects. The third is activity-level productivity, which is appropriate for specific activities. Based on construction management

perspective, productivity is often defined at the project level or activity level (Gerami Seresht and Fayek 2018). It is necessary to consider that the environment of CLP is unpredictable and complex, because of the large number of parameters that influence productivity directly or indirectly and the time-consuming process of tracking productivity. Therefore, providing a predictive model for productivity requires complex mapping of the multiple factors affecting labour productivity (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). In data mining, feature selection and extraction are necessary preprocessing approaches for identifying a relevant subset for classification and also developing a transformation of the inputs onto a low-dimensional subspace that keeps the majority of relevant information. The aim of using feature selection and extraction methods is to quickly develop prediction models with better performance.

Productivity prediction studies can be classified into three groups: statistical, simulation, and artificial intelligence (AI) techniques. The most common statistical technique is regression analysis. In general, regression models are limited by the number of influencing parameters and their capability of determining the combined impact of the influencing parameters (Song and AbouRizk 2008). System dynamics is one of the most applicable simulation techniques, and it is able to model a dynamic system. Although system dynamics models are able to capture the probabilistic uncertainties of real-world systems, they cannot capture the non-probabilistic uncertainties (i.e., subjective or linguistically expressed information) of real-world systems. On the other hand, AI techniques, such as artificial neural network (ANN), and their ability to learn from experience to improve their performance and adapt themselves to changes and also find patterns among datasets, make them useful methods for prediction (Mirahadi and Zayed 2016). For example, Gerami Seresht and Fayek (2018) developed the fuzzy system dynamics technique by integrating system dynamics and fuzzy logic to model multifactor productivity of equipment-intensive activities. Furthermore, Tsehayae and Fayek (2016a) demonstrated the application of data-driven fuzzy clustering in the development of fuzzy inference system (FIS). Then, they used a GA-based optimization process to address the FIS limitation, which is the inability to learn from data. Heravi and Eslamdoost (2015) developed an ANN model to predict CLP rates for foundation concrete work on industrial construction projects.

1.2 Problem Statement

CLP has been well studied because it has a direct effect on a company's efficiency and profitability and also because of the importance and vital role of labour productivity in improving project performance. The identification of factors that affect CLP is complex and vital for measuring and predicting construction productivity. Various studies have identified numerous factors influencing CLP, both subjective (e.g., foreman skill and task complexity) and objective (e.g., crew size). These studies have used questionnaire surveys to identify top factors influencing CLP (Alaghbari et al. 2019; Jarkas 2015; Montaser et al. 2018; Tsehayae and Fayek 2014).

Based on the above, **the first issue** in CLP modeling and prediction is related to identification of the most influential factors on CLP. Although several studies work on CLP factors identification, consensus on the classification and generalization of key parameters is yet to be achieved (Dixit et al. 2018). Additionally, the high-dimensional feature space of labour productivity often imposes a high computational cost as well as the risk of "overfitting" when classification is performed. Therefore, reducing the dimensionality of labour productivity data and identifying the most parameters that most influence CLP is necessary, since numerous factors have been identified that affect labour productivity. This can be done by using data mining techniques, namely feature selection methods. However, very few studies in CLP prediction use those methods to reduce the dimensionality of data. Filter and wrapper methods are the main approaches for feature selection (Yuan et al. 2018). 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. Furthermore, most filter methods are only suitable for developing mathematical equations by the statistical regression method (Gerami Seresht and Fayek 2018; Guyon et al. 2008). Wrapper methods, on the other hand, use the accuracy of a learning algorithm as a criterion for selecting useful features. 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 and Pedrycz 2012; Aličković and Subasi 2017). However, their application is limited because of the high computational complexity that occurs when numerous feature sets are considered. To resolve the afore-mentioned problem, it can be helpful to merge wrapper methods with suitable filter methods to reduce the wrapper method's deficiency, which is called hybrid feature selection (HFS). **The second issue** is related to the fact that most of past studies on modeling CLP used filter methods for selecting the most influential factors (Bai et

al. 2019; Gerami Seresht and Fayek 2018; Tsehayae and Fayek 2016b). The filter method is suitable for dealing with a high dimension of input space and a small number of data instances. However, using a wrapper method or HFS is more appropriate for predictive modeling using AI techniques, such as FIS and ANN, because of its superior performance (Piao and Ryu 2017).

The third issue is related to finding appropriate models for developing a predictive model and obtaining the optimal prediction evaluation index. Predicting CLP is still a challenge because of the limited CLP data availability to study, the complex variability of construction productivity, and the requirement of considering the complex effect of multiple variables simultaneously. However, maintaining high accuracy and interpretability in the developed models are the most important criteria. Predicting construction productivity has been accomplished in several studies mostly by using AI techniques (El-Gohary et al. 2017; Golnaraghi et al. 2020; Nasirzadeh et al. 2020; Sarihi et al. 2021). Despite the wide application of the predictive model of CLP for project planning and control, a predictive model on its sole application cannot offer the optimum point of the combination of influencing factors to construction companies for improving CLP. Concretely, no study has presented a hybrid model for finding the maximum value of CLP and optimum value of each influential factor using optimization techniques. Finding the maximum CLP helps project managers plan for improving each productivity factor. Therefore, **the fourth issue** in labour productivity modeling research is the absence of a general model to predict and optimize CLP in regards to finding the optimum value of each influential factor and the maximum value of CLP.

1.3 Research Objectives

The overall objective of this thesis is to develop a model for determining, predicting, and optimizing CLP at the activity level by using a combination of FS, AI, and evolutionary optimization techniques. To achieve this goal, this research has the following objectives:

1. Investigate the appropriate filter and wrapper methods as feature selection techniques for selecting the most value-adding factors of CLP and use the most effective combination of filter and wrapper methods as a hybrid method.
2. Predict CLP by applying and comparing different AI techniques, namely ANFIS, ANFIS-GA, RF, and ANN, using the most value-adding factors.
3. Develop an evolutionary optimization model for finding the maximum value of CLP by changing the value of each selected factor.

4. Develop a hybrid model to obtain a CLP value that is close to the company's preferred value of CLP and minimize deviation of predicted CLP factors from their average values in a dataset.

1.4 Expected Contributions

This thesis is expected to produce the following contributions that will positively impact future researchers and are classified under academic contributions, while some contributions will benefit industry practitioners and are classified under industrial contributions.

1.4.1 Academic contributions

The expected academic contributions of this research are:

- Development of a combination of filter and wrapper methods as an HFS to identify the most value-adding productivity factors and reduce the dimensionality of data
- Development of predictive models of CLP activity by evaluating different AI models' performance, namely ANFIS, ANFIS-GA, RF, and ANN using the selected factors
- Development of an evolutionary optimization model to obtain the maximum value of CLP and the optimum value of all the most value-adding factors

1.4.2 Industrial contributions

The expected industrial contributions of this research are:

- Identification of the most value-adding CLP factors, which helps construction planners provide improvement strategies and improve the most value-adding factors
- Prediction of labour productivity for use in construction project cost estimation and scheduling
- Development of a hybrid model for optimizing CLP and its factors, which will be effective for construction planners to carry out productivity improvement studies and analyze different scenarios

1.5 Research Methodology

The objectives of this research are achieved in five main stages, which are listed below.

1.5.1 The first stage: Identification of factors influencing CLP

The development of productivity modeling and optimization begins with the identification of the factors influencing productivity. By analyzing existing literature in the field of CLP analysis and modeling, the factors influencing productivity are identified. Finally, identification of the most appropriate feature selection techniques is reviewed.

1.5.3 The second stage: Feature selection

The large input parameters feature space, made up of the influencing factors, had to be reduced in order to maintain the interpretability and accuracy of the productivity models. To overcome these challenges and find the factors with the most influence on CLP, feature selection methods are used. The feature space is reduced by identifying the key input factors influencing productivity using feature selection methods. The HFS, which is the integration of filter and wrapper method, is developed to find the factors most influencing CLP.

1.5.4 The third stage: The predictive model development

The predictive model for CLP is developed by using the selected features as inputs and CLP as an output. Different AI techniques, namely ANFIS, ANFIS-GA, RF, and ANN, are applied for developing three different predictive models. As shown in the current literature, ANN has become a popular and helpful model for classification, clustering, pattern recognition, and prediction in many disciplines. One advantage of ANN is the high-speed processing provided in a massive parallel implementation. ANNs are able to deal with noisy or incomplete data and can be very effective, especially in problems where the relationships between inputs and outputs are not sufficiently known (Almási et al. 2016). So, based on their abilities, ANNs can be ideal alternatives for modeling labour productivity. ANFIS is one of the most popular neuro-fuzzy systems. In ANFIS, learning ability and relational structure of the ANNs is combined with the decision-making mechanism of fuzzy logic (Siraj et al. 2016). In order to optimize ANFIS parameters, the integration of ANFIS and GA is also developed. Another algorithm that shows accurate performance in a number of studies in other disciplines is RF. Comparing the predicted results of ANN as an AI technique, ANFIS as a neuro-fuzzy system, and RF as a classifier, the performance

evaluation of these models based on the root mean square error (RMSE) and accuracy of predicted results are achieved, and the most accurate model is selected for prediction of CLP.

1.5.5 The fourth stage: Evolutionary optimization model development

A hybrid model using particle swarm optimization (PSO) is developed to find the maximum amount of CLP and optimum value of each factor. Due to the limitations of mathematical optimizations such as reaching local optimum, evolutionary-based algorithms have been proposed for finding a near-optimal solution space with better results.

1.6 Thesis Organization

Chapter 1 provides background information on CLP research and identifies gaps in the CLP research. This chapter also presents the research objectives, expected academic and industrial contributions, and research methodology of the thesis.

Chapter 2 presents an extensive literature review on the relevant topics, including feature selection methods, identification of factors influencing CLP, and development of predictive models for CLP.

Chapter 3 presents the methodology of the proposed hybrid model, which contains (1) CLP dataset overview, (2) data preparation, (3) HFS process, (4) CLP predictive modeling, and (5) CLP optimization.

Chapter 4 presents the application of the developed hybrid model for CLP prediction and optimization. This chapter focuses on experimental results and discussion regarding the proposed model.

Chapter 5 describes the conclusions, contributions, and limitations of the thesis, as well as recommendations for future research.

1.7. References

Ahmad, S. S. S., and Pedrycz, W. (2012). "Data and feature reduction in fuzzy modeling through particle swarm optimization." *Applied Computational Intelligence and Soft Computing*, 2012(Article ID 347157), 1–21.

Alaghbari, W., Al-Sakkaf, A. A., and Sultan, B. (2019). "Factors affecting construction labour productivity in Yemen." *International Journal of Construction Management*, 19(1), 79–91.

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

Almási, A. D., Woźniak, S., Cristea, V., Leblebici, Y., and Engbersen, T. (2016). "Review of advances in neural networks: Neural design technology stack." *Neurocomputing*, 174, 31–41.

Bai, S., Li, M., Kong, R., Han, S., Li, H., and Qin, L. (2019). "Data mining approach to construction productivity prediction for cutter suction dredgers." *Automation in Construction*, 105, 102833.

Dixit, S., Mandal, S. N., Thanikal, J. V, and Saurabh, K. (2018). "Construction productivity and construction project performance in Indian construction projects."*Proceedings Creative Construction Conference 2018*, 379–386. Diamond Congress Ltd., Budapest University of Technology and Economics.

Doloi, H. (2008). "Application of AHP in improving construction productivity from a management perspective." *Construction Management and Economics*, 26(8), 841–854.

Eastman, C. M., and Sacks, R. (2008). "Relative productivity in the AEC industries in the United States for on-site and off-site activities." *Journal of Construction Engineering and Management* (ASCE), 134(7), 517–526.

El-Gohary, K. M., Aziz, R. F., and Abdel-Khalek, H. A. (2017). "Engineering approach using ANN to improve and predict construction labour productivity under different influences." *Journal of Construction Engineering and Management*, 143(8), 04017045.

Gerami Seresht, N., and Fayek, A. R. (2018). "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management* (ASCE), 144(9), 04018091.

Golnaraghi, S., Moselhi, O., Alkass, S., and Zangenehmadar, Z. (2020). "Predicting construction labour productivity using lower upper decomposition radial base function neural network." *Engineering Reports*, 2(2), 1–16.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., eds. (2006). *Feature extraction: Foundations and applications*. Berlin/Heidelberg, Germany: Springer-Verlag.

Heravi, G., and Eslamdoost, E. (2015). "Applying artificial neural networks for measuring and predicting construction-labour productivity." *Journal of Construction Engineering and Management*, 141(10), 04015032.

Jarkas, A. M. (2015). "Factors influencing labour productivity in Bahrain's construction industry." *International Journal of Construction Management*, 15(1), 94–108.

Loosemore, M. (2014). "Improving construction productivity: a subcontractor's perspective." *Engineering, Construction and Architectural Management*, 21(3), 245–260.

Mirahadi, F., and Zayed, T. (2016). "Simulation-based construction productivity forecast using eural-network–driven fuzzy reasoning." *Automation in Construction*, 65, 102–115.

Montaser, N. M., Mahdi, I. M., Mahdi, H. A., and Rashid, I. A. (2018). "Factors affecting construction labour productivity for construction of pre-stressed concrete bridges." *International Journal of Construction Engineering and Management*, 7(6), 193–206.

Nasirzadeh, F., Kabir, H. M. D., Akbari, M., Khosravi, A., Nahavandi, S., and Carmichael, D. G. (2020). "ANN-based prediction intervals to forecast labour productivity." *Engineering, Construction and Architectural Management*, 27(9), 2335–2351.

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." *Asian Conference on Intelligent Information and Database Systems*, 721–727. Cham, Switzerland: Springer.

Sarihi, M., Shahhosseini, V., and Banki, M. T. (2021). "Development and comparative analysis of the fuzzy inference system-based construction labour productivity models." *International Journal of Construction Management*, 0(0), 1–18.

Siraj, N. B., Fayek, A. R., and Tsehayae, A. A. (2016). "Development and optimization of artificial intelligence-based concrete compressive strength predictive models." *International Journal of Structural and Civil Engineering Research*, 5(3), 156–167.

Song, L., and Abourizk, S. M. (2008). "Measuring and modeling labour productivity using historical data." *Journal of Construction Engineering and Management*, 134(10), 786–794.

Statistics Canada. (2019). Gross domestic product at basic prices, by industry, 2015–2019. <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610043403&pickMembers%5B0%5D=2.1&pickMembers%5B1%5D=3.1&cubeTimeFrame.startYear=2016&cubeTimeFrame.endYear=2020&referencePeriods=20160101%2C20200101 >.

Talhouni, B. (1990). *Measurement and analysis of construction labour productivity* (Doctoral dissertation). University of Dundee, Dundee, Scotland.

Tsehayae, A. A., and Fayek, A. R. (2014). "Identification and comparative analysis of key parameters influencing construction labour productivity in building and industrial projects." *Canadian Journal of Civil Engineering*, 41(10), 878–891.

Tsehayae, A. A., and Fayek, A. R. (2016a). "Developing and optimizing context-specific fuzzy inference system-based construction labour productivity models." *Journal of Construction Engineering and Management*, 142(7), 04016017.

Tsehayae, A. A., and Fayek, A. R. (2016b). "System model for analysing construction labour productivity." *Construction Innovation*, 16(2), 203–228.

Yuan, H., Xu, G., Yao, Z., Jia, J., and Zhang, Y. (2018). "Imputation of missing data intime series for air pollutants using long short-term memory recurrent neural networks." *Proceedings 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 1293–1300.

Chapter 2. Literature Review

2.1. Feature Selection Methods

Feature selection is a vital part of any data mining process, which reduces the number of features by selecting a subset of the input features that efficiently represents the input data while removing irrelevant, noisy, and redundant data, and results in acceptable classification accuracy (Hall 1999; Bai et al. 2019). The main benefits of implementing feature selection methods 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). Feature selection methods can be broadly categorized as filter, wrapper, and embedded methods (Wei et al. 2020).

Filter methods offer less computational time to provide results than wrapper or embedded methods (Atallah et al. 2019). They do not require any learning algorithm; rather, they evaluate statistical properties and rank features based on heuristic scoring criteria. As filter methods are independent of classifiers, selected features are not the most appropriate for all classifiers (Ghosh et al. 2019; Lee et al. 2017). Although filter methods are fast and computationally less expensive, their accuracy is low, since the selection process does not involve any classifiers. Filter methods can determine optimum features as a subset by using various measures, namely distance, information (or uncertainty), dependence, and consistency measures (Dash and Liu 1997). A brief explanation of these measures is provided as follows.

1) Distance measure is also known as divergence or discrimination measurement. Generally, this type of measure is primarily utilized for two-class problems, although it can be expanded for multi-class problems as well. Various methods can be used for difference measurement, such as Euclidean and Manhattan. Relief method is one of the most well-known techniques of this category. Relief feature scoring is based on the identification of distance between nearest neighbor instance pairs.
2) Information measure usually computes the entropy or uncertainty of features based on class labels and thereby calculates information gain. As an example, in a feature problem, the feature M is more preferable than another feature N if the information gain of M is more than that of N. Information gain of a feature M is specified by the difference between the

prior uncertainty of feature *M* and the expected posterior uncertainty using that feature. Mutual information is a popular technique in this category that quantifies the amount of information achieved from a random feature, through the other random features.

3) Dependence measure is also known as a correlation measurement for predicting the value of one factor based on that of another factor. By utilizing several classical dependence measures, the correlation between a feature and a class is obtained. If the correlation value between the feature *M* and class *L* is higher than feature *N* and class *L*, then feature *M* is more preferable than *N*. Chi-square feature ranking, which belongs to this category, measures the degree of association between two categorical features.
4) Consistency measure relies on the training dataset and the use of min-features bias in selecting a subset of features. Min-feature bias identifies the minimally sized subset that satisfies the acceptable rate of inconsistency, which is set by the user (Dash and Liu 1997).

Wrapper methods use the accuracy of a learning algorithm (e.g., classification algorithm) as a criterion for selecting features (Venkatesh and Anuradha 2019). Thus, wrapper methods are tuned to the specific interaction between a learning algorithm and its training data. Recently, wrapper methods have received a lot of attention because of their better generalization performance compared to filter methods (Mafarja and Mirjalili 2018; Monirul Kabir et al. 2010). A wrapper method starts with a given feature subset (which can be selected randomly) and evaluates each generated feature subset by applying a learning model to the dataset. If the performance of the generated feature subset improves, it is selected as a current best feature subset. However, wrapper method applications are limited by the high computational complexity when feature sets are wide and also a risk of overfitting (Monirul Kabir et al. 2010; Piao and Ryu 2017).

Embedded methods simultaneously determine features and classifiers during the model training process. These methods have the merit of interacting with learning algorithms to perform feature selection during the model training. However, for high-dimensional data, embedded methods lead to computationally expensive calculations (Liu et al. 2019; Wei et al. 2020).

In recent years, many feature selection algorithms have been proposed in order to overcome limitations of the standard methods (e.g., filter, wrapper, and embedded) in different datasets. HFS methods as a combination of filter and wrapper methods are exploited to achieve better accuracy than filter methods and reduce computation and complexity of wrapper methods (Nguyen et al.

2020). The general approach for hybrid methods is a feature selection method consisting of two stages. In the first stage, a filter method refines features (mostly a ranking technique selects the top-*n* features), and in the second stage a wrapper method finds the most discriminative subset from the top-*n* features (Ghosh et al. 2019). Different HFS methods have been proposed by researchers. Lee and Leu (2011) proposed a novel HFS method for feature selection in microarray data analysis by using a genetic algorithm (GA) with dynamic parameter setting (GADP) to generate a number of subsets of genes and rank the genes according to their occurrence frequencies in the gene subsets. They used the χ^2 test as a feature ranking method to select a proper number of top-ranked features and concluded that the proposed GADP method selects fewer genes while giving higher prediction accuracy. Hsu et al. (2011) introduced another HFS method that used F-score and information gain (IG) as filter methods to remove the most redundant or irrelevant features, and they applied support vector machine (SVM) as a wrapper method to those selected features for further data reduction. Kari et al. (2018) proposed a hybrid approach combined with a GA and SVM to improve fault diagnosis accuracy in power transformers. Fei and Min (2016) presented a novel optimization approach by using SVM to select a support vector subset and feature subset simultaneously based on GA to solve binary classification. Tao et al. (2019) proposed an approach of feature selection and parameter optimization of SVM using GA for hospitalization expense modeling, which includes binary data sets. Lu et al. (2017) introduced the combination of mutual information maximization and adaptive GA as a new HFS method. The experimental results of these HFS methods indicate that the proposed methods have the ability to reduce time complexity and improve classification accuracy. Therefore, HFS methods are capable of obtaining relevant features from large datasets with a high classification accuracy.

Because numerous factors affect CLP and cause complexity in predicting and modeling labour productivity, HFS methods can help obtain relevant features from a complex CLP dataset and select an adequate data sample for subsequent CLP analysis. Previous studies mostly focus on using questionnaires and determining some statistical measures such as relative importance index (RII) and mean response (MR) to identify the most influential CLP and are reviewed in the following section.

2.2. CLP Measurement and Influencing Factors

CLP is affected by numerous variables made up of various objective and subjective factors (e.g., crew size, complexity of task, and weather condition). Previous studies used questionnaire surveys to identify top factors influencing CLP (Agrawal and Halder 2020; Alaghbari et al. 2019; Durdyev et al. 2018; Irfan et al. 2020; Jarkas 2015; Montaser et al. 2018; Tsehayae and Fayek 2014). Table 2.1 shows a brief description of several studies that have identified top factors influencing CLP using statistical analysis such as RII, MR, and frequency index.

Table 2.1. Summary of previous research on CLP factors identification

Source	Methods	Most influencing factors on CLP
Hafez (2014)	Using a questionnaire survey comprising 27 productivity factors, identified and ranked them using RII measure	Top 7 factors influencing CLP: (1) payment delay, (2) skills of labour, (3) shortage of experienced labour, (4) lack of labour supervision, (5) motivation of labour, (6) working overtime, (7) lack of leadership of construction managers
Chigara and Moyo (2014)	Using a questionnaire, which included 40 preselected CLP factors, which were ranked using RII and MR measures	Top 5 factors influencing CLP: (1) materials unavailability, (2) late payment of salaries, (3) plant and equipment suitability/adequacy, (4) supervisory incompetence, (5) lack of manpower/skills
Tsehayae and Fayek (2014)	Investigating the influence of 169 parameters on CLP using project management and trade surveys; positive and negative influences of CLP factors evaluated using statistical analysis to identify top CLP factors	Top 5 factors influencing CLP: (1) adequate and quality work tools, (2) aging of Canada's population, (3) job site orientation program for new craftsmen, (4) lack of protection from weather effect, (5) use of daily job assessment system

Jarkas (2015)	Utilizing a structured questionnaire survey, which found 37 labour productivity factors, and using RII to identify the factors most influential on CLP	Top 6 factors influencing CLP: (1) skills of labour, (2) design disciplines coordination, (3) lack of labour supervision, (4) design drawings errors and omissions, (5) delay in responding to requests for information, (6) rework
Montaser et al. (2018)	A total of 50 respondents consisting of owners, contractors, and consultants were asked to indicate the importance of CLP factors; top CLP factors selected using importance index and frequency index	Top 5 factors influencing CLP: (1) lack of structure system and design cables, (2) absence of authority to discipline labour, (3) rework in drawing, (4) lack of communication between workers and engineers, (5) slow response of the consultant
Alaghbari et al. (2019)	Using a questionnaire comprising 52 predefined factors and identifying the most influencing factors on CLP from the perspective of structural engineers by determining RII technique	Top 5 factors influencing CLP: (1) experience and skills of labourers, (2) availability of materials in site, (3) leadership and efficiency in site management, (4) availability of materials in the market, (5) political and security situation
Agrawal and Halder (2020)	Using a structured questionnaire survey comprising 29 labour productivity factors; identified factors ranked using RII technique	Top 7 factors influencing CLP: (1) labour personal problems, (2) improper managerial skills, (3) scheduling of work, (4) high or low temperature, (5) schedule compression, (6) labourer dissatisfaction, (7) shortage of materials

Most previous studies used questionnaire surveys to identify the most influential CLP factors. However, the selected factors highly relied on expert knowledge, which can be very changeable from time to time. On the other hand, identification of the factors most influencing CLP without using questionnaires is a challenging process because of the limited historical data on CLP influential factors. Several studies in labour productivity used filter feature selection methods to identify top CLP factors. Tsehayae and Fayek (2016) used a correlation-based feature selection (CFS), which is a filter method, to find the key influencing CLP features. The CFS algorithm is appropriate because of its ability to deal with a high-dimensional feature space. However, wrapper

or HFS methods are more appropriate for predictive modeling that uses AI techniques, such as FISs, ANNs, and SVMs, because of their superior performance based on learning algorithms (Piao and Ryu 2017). Several studies showed that the use of wrapper or HFS method in the application, where the predictive model is developed, shows better results for accuracy (Ahmad and Pedrycz 2012; Gerami Seresht et al. 2020).

2.3. Modeling CLP

Modeling CLP is challenging because the impact of numerous factors must be considered simultaneously. A series of modeling techniques, such as regression models, system dynamics, and ANN, have been introduced to map the relationship between CLP and factors influencing it (El-Gohary et al. 2017). Regression is one of the most common modeling techniques for CLP (Song and Abourizk 2008). Thomas and Sudhakumar (2014) developed several linear regression models to determine the effect of 11 factors influencing masonry labour productivity. Parthasarathy et al. (2018) developed 15 different models using multiple linear regression analysis to model manpower and equipment productivity in tall residential building projects. Hai and Tam (2020) presented a multiple linear regression model as a statistical method for output prediction, to evaluate the impact of 10 factor groups on CLP. However, regression models have a number of limitations, such as lack of capacity to deal with large number of inputs and intolerance to noisy data (Lu 2000; Tsehayae 2015). Also, multiple linear regression analysis needs each input factor to contain a linear relationship with CLP. Since CLP factors are often related to each other, the multiple linear regression analysis does not cover the requirements for CLP modeling.

To deal with the aforementioned challenges, AI techniques such as FISs, ANNs, decision tree classifiers, and SVMs are widely used in the construction management domain (Cheng et al. 2021). Golnaraghi et al. (2020) developed a prediction model for CLP by using ANN and compared them with other techniques including ANFIS and radial basis function neural network. The results showed the superior performance of radial basis function neural network compared to other models. El-Gohary et al. (2017) introduced the engineering approach using ANN techniques to map the relationship between CLP and its influential factors. Nasirzadeh et al. (2020) developed ANN-based prediction intervals to predict CLP using historical data. Their model provided various sources of uncertainty affecting prediction. Momade et al. (2020) proposed a data-driven approach

using SVM and RF to model and predict CLP. Their results showed the SVM model achieved higher rate of accuracy compared to RF.

However, in recent years, the hybrid systems based machine learning, optimization algorithm and simulation techniques have been applied in several construction problems due to their superiority over sole AI techniques (Cheng et al. 2020; Zhang et al. 2020). Gerami Seresht and Fayek (2018), developed the fuzzy system dynamics technique by integrating system dynamics and fuzzy logic to model multifactor productivity of equipment-intensive activities. Furthermore, Tsehayae and Fayek (2016), demonstrated the application of a data-driven fuzzy clustering in the development of FIS. Then, they used a GA-based optimization process to address the FIS limitation which is the inability to learn from data. Khanzadi et al. (2017) developed a hybrid simulation model by combining system dynamics and agent-based modeling to predict labour productivity by considering various influencing factors in a concreting project. Raoufi and Fayek (2018) proposed the integration of fuzzy logic and agent-based modeling to predict the performance of construction crews according to crew motivational and situational input variables. Gerami Seresht et al. (2020) introduced a new fuzzy clustering algorithm by using Gustafson-Kessel's algorithm and Adam optimization method to determines the number of clusters automatically and assigns weights to the FIS rules to improve accuracy. Then, the proposed algorithm was used to predict CLP for concrete placing activities and the results showed the new approach improves the accuracy and efficiency compared to the past research.

Although the aforementioned papers developed the hybrid methods to model and predict construction productivity, very few studies applied HFS methods, as the combination of filter and wrapper feature selection methods, to construction productivity prediction to reduce dimensionality and to find the most predictive factors. Ebrahimi et al. (2020), proposed the integration of ANN and GA as a wrapper method for feature selection and predicting CLP. The results showed an improvement in accuracy compared to previous works using filter methods. Recently, Cheng et al. (2021) introduced the hybrid model including least square SVM, symbiotic organisms search, and wrapper-based feature selection methods to predict construction productivity. This thesis presents an HFS method to identify the most predictive CLP factors to utilize them as inputs of the developed CLP predictive models.

Despite wide application of the predictive model of CLP for project planning and control, a predictive model in its sole application cannot offer construction companies the optimum value of influencing factors for improving CLP. In the construction domain, most optimization studies were found in the context of optimizing time-cost trade-off models. Lin and Lai (2020) introduced an optimized time-cost trade-off model that considers variable productivity related to working environment and management. They applied GA to identify optimal and near-optimal labour productivity. Dehghan et al. (2015) presented an overlapping optimization algorithm based on GA principles to develop a practical approach to determining optimal overlapping activities in construction projects. Very few studies focus on finding the optimal CLP in construction projects. Kisi et al. (2017) introduced a two-prong strategy for estimating optimal productivity in labour-intensive construction operations. The first prong estimates the upper limit of optimal productivity by using a qualitative factor model. The second prong estimates the lower limit of the productivity by removing operational inefficiencies from actual value of productivity by using a discrete-event simulation. However, hybrid optimization, which simultaneously predicts and optimizes CLP while considering its influential factors, has not been explored in the area of labour productivity optimization. Therefore, this thesis develops a hybrid evolutionary optimization technique by integrating HFS, a predictive model, and an evolutionary optimization technique to optimize CLP and its influential factors as a novel method to find the maximum value of CLP while considering minimum changes to CLP key factors.

2.4. Summary

This chapter provides a literature review on the feature selection methods, factors influencing CLP, and different techniques of modeling CLP in previous studies and identifies the research gaps in these topics. The existing gaps in the CLP literature include: (1) lack of research on identification of the most influential CLP factors using HFS methods as an important preprocessing procedure for data mining, and (2) lack of research on developing a hybrid CLP model for prediction and optimization using its most influential factors. There are many studies on the identification of factors most influencing CLP. However, most previous studies highly relied on expert knowledge and statistical analysis without any learning algorithm, which can be a limitation for developing a CLP predictive model using AI techniques. There are also very few studies that use HFS methods to identify the factors most influencing CLP. Furthermore, CLP prediction and optimization are still challenging because of the limited CLP data availability to study, the complex variability of

construction productivity, and the modeling requirement of simultaneously considering the complex effect of multiple variables. The next chapter presents the methodology for identifying the factors that most influence CLP.

2.5. References

Agrawal, A., and Halder, S. (2020). "Identifying factors affecting construction labour productivity in India and measures to improve productivity." *Asian Journal of Civil Engineering*, 21(4), 569–579.

Ahmad, S. S. S., and Pedrycz, W. (2012). "Data and feature reduction in fuzzy modeling through particle swarm optimization." *Applied Computational Intelligence and Soft Computing*, 2012(Article ID 347157), 1–21.

Alaghbari, W., Al-Sakkaf, A. A., and Sultan, B. (2019). "Factors affecting construction labour productivity in Yemen." *International Journal of Construction Management*, 19(1), 79–91.

Atallah, D. M., Badawy, M., El-Sayed, A., and Ghoneim, M. A. (2019). "Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier." *Multimedia Tools and Applications*, 78(14), 20383–20407.

Bai, S., Li, M., Kong, R., Han, S., Li, H., and Qin, L. (2019). "Data mining approach to construction productivity prediction for cutter suction dredgers." *Automation in Construction*, 105, 102833.

Cheng, M.-Y., Cao, M.-T., and Jaya Mendrofa, A. Y. (2020). "Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine." *Journal of Building Engineering*, 101973.

Cheng, M. Y., Cao, M. T., and Jaya Mendrofa, A. Y. (2021). "Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine." *Journal of Building Engineering*, 35, 101973.

Chigara, B., and Moyo, T. (2014). "Factors affecting labour productivity on building projects in Zimbabwe." *International Journal of Architecture, Engineering and Construction*, 3(1), 57–65.

Dash, M., and Liu, H. (1997). "Feature selection for classification." *Intelligent Data Analysis*, 1(3), 131–156.

Dehghan, R., Hazini, K., and Ruwanpura, J. (2015). "Optimization of overlapping activities in the design phase of construction projects." *Automation in Construction*, 59, 81–95.

Durdyev, S., Ismail, S., and Kandymov, N. (2018). "Structural equation model of the factors affecting construction labour productivity." *Journal of Construction Engineering and Management*, 144(4), 04018007.

Ebrahimi, S., Raoufi, M., and Fayek, A. R. (2020). "Framework for integrating an artificial neural network and a genetic algorithm to develop a predictive model for construction labour productivity." *Construction Research Congress 2020*, American Society of Civil Engineers, Reston, VA, 58–66.

El-Gohary, K. M., Aziz, R. F., and Abdel-Khalek, H. A. (2017). "Engineering approach using ANN to improve and predict construction labour productivity under different influences." *Journal of Construction Engineering and Management*, 143(8), 04017045.

Fei, Y., and Min, H. (2016). "Simultaneous feature with support vector selection and parameters optimization using GA-based SVM solve the binary classification." *2016 1st IEEE International Conference on Computer Communication and the Internet, ICCCI 2016*, 426–433.

Gerami Seresht, N., and Fayek, A. R. (2018). "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management* (ASCE), 144(9), 04018091.

Gerami Seresht, N., Lourenzutti, R., and Fayek, A. R. (2020). "A fuzzy clustering algorithm for developing predictive models in construction applications." *Applied Soft Computing*, 96, 106679.

Ghosh, M., Guha, R., Sarkar, R., and Abraham, A. (2019). "A wrapper-filter feature selection technique based on ant colony optimization." *Neural Computing and Applications*, 32, 7839–7857.

Golnaraghi, S., Moselhi, O., Alkass, S., and Zangenehmadar, Z. (2020). "Predicting construction labour productivity using lower upper decomposition radial base function neural network." *Engineering Reports*, 2(2), 1–16.

Hafez, S. M. (2014). "Critical factors affecting construction labour productivity in Egypt." *American Journal of Civil Engineering*, 2(2), 35.

Hai, D. T., and Van Tam, N. (2020). "Application of the regression model for evaluating factors affecting construction workers' labour productivity in Vietnam." *The Open Construction and Building Technology Journal*, 13(1), 353–362.

Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation). University of Waikato, Hamilton, New Zealand.

Hsu, H. H., Hsieh, C. W., and Lu, M. Da. (2011). "Hybrid feature selection by combining filters and wrappers." *Expert Systems with Applications*, 38(7), 8144–8150.

Irfan, M., Zahoor, H., Abbas, M., and Ali, Y. (2020). "Determinants of labour productivity for building projects in Pakistan." *Journal of Construction Engineering, Management & Innovation*, 3(2), 85–100.

Jarkas, A. M. (2015). "Factors influencing labour productivity in Bahrain's construction industry." *International Journal of Construction Management*, 15(1), 94–108.

Kari, T., Gao, W., Zhao, D., Abiderexiti, K., Mo, W., Wang, Y., and Luan, L. (2018). "Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm." *IET Generation, Transmission and Distribution*, 12(21), 5672–5680.

Khanzadi, M., Nasirzadeh, F., Mir, M., and Nojedehi, P. (2017). "Prediction and improvement of labour productivity using hybrid system dynamics and agent-based modeling approach." *Construction Innovation*, 18(1), 2–19.

Kisi, K. P., Mani, N., Rojas, E. M., and Foster, E. T. (2017). "Optimal productivity in labour-intensive construction operations: Pilot study." *Journal of Construction Engineering and Management*, 143(3), 04016107.

Lee, C. P., and Leu, Y. (2011). "A novel hybrid feature selection method for microarray data analysis." *Applied Soft Computing*, 11(1), 208–213.

Lee, J., Park, Y. J., Choi, C. H., and Han, C. H. (2017). "BIM-assisted labour productivity measurement method for structural formwork." *Automation in Construction*, 84, 121–132.

Lin, C. L., and Lai, Y. C. (2020). "An improved time-cost trade-off model with optimal labour

productivity." *Journal of Civil Engineering and Management*, 26(2), 113–130.

Liu, H., Zhou, M., and Liu, Q. (2019). "An embedded feature selection method for imbalanced data classification." *IEEE/CAA Journal of Automatica Sinica*, 6(3), 703–715.

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., and Gao, Z. (2017). "A hybrid feature selection algorithm for gene expression data classification." *Neurocomputing*, 256, 56–62.

Lu, M. (2000). *Productivity studies using advanced ANN models* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses (204).

Mafarja, M., and Mirjalili, S. (2018). "Whale optimization approaches for wrapper feature selection." *Applied Soft Computing*, 62, 441–453.

Momade, M. H., Shahid, S., Hainin, M. R. bin, Nashwan, M. S., and Tahir Umar, A. (2020). "Modelling labour productivity using SVM and RF: a comparative study on classifiers performance." *International Journal of Construction Management*, 0(0), 1–11.

Monirul Kabir, M., Monirul Islam, M., and Murase, K. (2010). "A new wrapper feature selection approach using neural network." *Neurocomputing*, 73(16–18), 3273–3283.

Montaser, N. M., Mahdi, I. M., Mahdi, H. A., and Rashid, I. A. (2018). "Factors affecting construction labour productivity for construction of pre-stressed concrete bridges." *International Journal of Construction Engineering and Management*, 7(6), 193–206.

Nasirzadeh, F., Kabir, H. M. D., Akbari, M., Khosravi, A., Nahavandi, S., and Carmichael, D. G. (2020). "ANN-based prediction intervals to forecast labour productivity." *Engineering, Construction and Architectural Management*, 27(9), 2335–2351.

Nguyen, B. H., Xue, B., and Zhang, M. (2020). "A survey on swarm intelligence approaches to feature selection in data mining." *Swarm and Evolutionary Computation*, 54(October 2019), 100663.

Parthasarathy, M. K., Murugasan, R., and Vasan, R. (2018). "Modelling manpower and equipment productivity in tall residential building projects in developing countries." *Journal of the South African Institution of Civil Engineering*, 60(2), 23–33.

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." *Proceedings Asian Conference on Intelligent Information and Database Systems*, 721–727. Cham, Switzerland: Springer.

Raoufi, M., and Fayek, A.R. (2018). "Fuzzy agent-based modeling of construction crew motivation and performance." *Journal of Computing in Civil Engineering*, 32(5), 04018035.

Song, L., and Abourizk, S. M. (2008). "Measuring and modeling labour productivity using historical data." *Journal of Construction Engineering and Management*, 134(10), 786–794.

Tao, Z., Huiling, L., Wenwen, W., and Xia, Y. (2019). "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling." *Applied Soft Computing*, 75, 323–332.

Thomas, A. V., and Sudhakumar, J. (2014). "Modelling masonry labour productivity using multiple regression." *Proceedings 30th Annual Association of Researchers in Construction Management Conference, ARCOM 2014*, 1345–1354.

Tsehayae, A. (2015). *Developing and optimizing context-specific and universal construction labour productivity models* (Doctoral dissertation). University of Alberta, Edmonton, Alberta.

Tsehayae, A. A., and Fayek, A. R. (2014). "Identification and comparative analysis of key parameters influencing construction labour productivity in building and industrial projects." *Canadian Journal of Civil Engineering*, 41(10), 878–891.

Tsehayae, A. A., and Fayek, A. R. (2016). "Developing and optimizing context-specific fuzzy inference system-based construction labour productivity models." *Journal of Construction Engineering and Management*, 142(7), 04016017.

Venkatesh, B., and Anuradha, J. (2019). "A hybrid feature selection approach for handling a high-dimensional data." *Proceedings of the 6th International Innovations in Computer Science and Engineering Conference*, edited by Saini H., Sayal R., Govardhan A., and Buyya R., pages 365–373. Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, volume 74. Springer: Singapore.

Wei, G., Zhao, J., Feng, Y., He, A., and Yu, J. (2020). "A novel hybrid feature selection method

based on dynamic feature importance." *Applied Soft Computing*, 93, 106337.

Zhang, J., Li, D., and Wang, Y. (2020). "Predicting uniaxial compressive strength of oil palm shell concrete using a hybrid artificial intelligence model." *Journal of Building Engineering*, 30, 101282.

Chapter 3. Development of the hybrid model for CLP prediction and optimization[1, 2]

3.1. Introduction

The accurate prediction of CLP is vital since it helps construction managers avoid cost overrun and falling behind schedule (Grau et al. 2009). Accordingly, several AI techniques have been successfully applied to modeling and predicting construction productivity, which is discussed in the literature review section. CLP is affected by numerous factors that reduce the accuracy of the predictive model and impose the risk of data overfitting (Ebrahimi et al. 2020). Feature selection is one of the important preprocessing procedures for data mining. Therefore, it is necessary to apply effective feature selection methods that are able to select key features affecting CLP and reject the nonessential features in order to achieve high prediction accuracy and reduce model complexity (Atallah et al. 2019; Topuz et al. 2018). HFS methods are a combination of filter and wrapper methods and therefore reduce deficiencies of both methods (Venkatesh and Anuradha 2019; Piao and Ryu 2017). Although comprehensive studies have identified CLP factors, few works have focused on applying different feature selection methods to CLP factors to reduce the risk of overfitting. In other words, a research gap exists regarding development of HFS methods as an essential data cleaning process prior to CLP modeling.

Despite wide application of predictive CLP models for project planning and control, a predictive model as a sole application cannot offer construction companies the optimum value of influencing factors for improving CLP (Cheng et al. 2021). Concretely, no study has presented a hybrid model for finding the maximum value of CLP and optimum value of each influential factor using optimization techniques. Although, there are many studies in CLP prediction, their main limitation is the lack of capacity to optimize CLP and its most predictive factors with respect to a construction company's preferences, such as a targeted CLP.

---

[1] Parts of this chapter have been accepted for publication: Ebrahimi, S., Fayek, A. R., and Sumati, V. (2021). "Hybrid artificial intelligence HFS-RF-PSO model for construction labor productivity prediction and optimization." *Algorithms* 14(7), 214.

[2] Parts of this chapter have been submitted for publication Ebrahimi, S.; Kazerooni, M.; Sumati, V.; Fayek, A. R. (n.d.). "A predictive model for construction labour productivity using the integration of hybrid feature selection and PCA methods." *Canadian Journal of Civil Engineering*, under review.

This thesis aims to fill the gap in the literature by developing a hybrid model that can identify the factors that most influence CLP as well as predict and optimize CLP and the factors influencing it. The proposed hybrid model will help project managers have more confidence in predicted CLP and be able to plan for improving each CLP factor.

The major contribution of this thesis is developing a model for both predicting and optimizing CLP using a combination of feature selection, AI, and evolutionary optimization techniques. To achieve this goal, this thesis had the following objectives: (1) identify factors that are most predictive of CLP using a combination of filter and wrapper methods as an HFS method, (2) predict CLP by developing and comparing four different predictive models using the factors that most influence CLP, and (3) develop a novel hybrid evolutionary optimization model for finding the maximum CLP value and the optimum value of each selected factor.

3.2. Model Development

This section discusses the methodology of a hybrid model for CLP prediction, and optimization, which consists of five steps: (1) CLP dataset overview, (2) CLP data preparation, (3) HFS technique, (4) CLP predictive models development, and (5) CLP optimization. Figure 3.1 illustrates these five steps, which are further discussed in the following sub-sections.

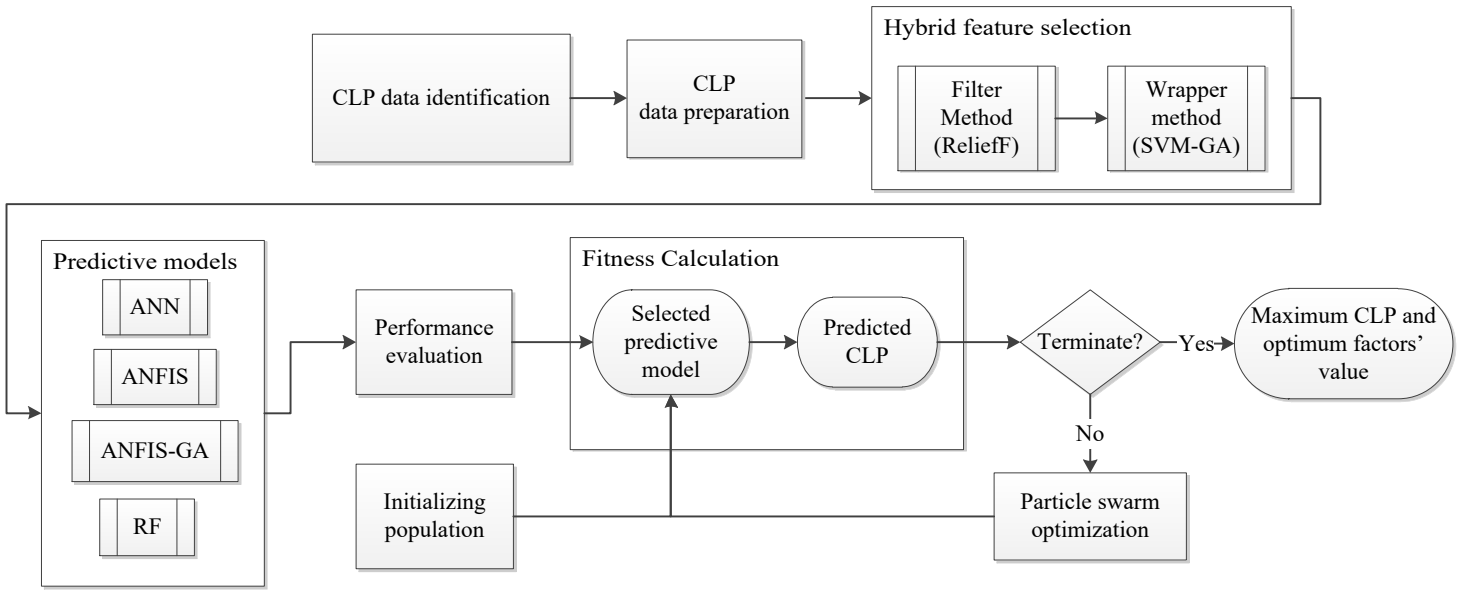

Figure 3.1. A general view of the proposed research methodology for CLP prediction and optimization

3.2.1. CLP dataset overview

In this thesis, the proposed methodology was used to predict and optimize CLP of concrete placing activities, using the data collected by Tsehayae and Fayek (2016a) in a previous study. Data were collected in Alberta, Canada, in four construction project contexts, including residential and commercial warehouse buildings, residential and commercial high-rise buildings, industrial buildings, and institutional buildings. A literature review conducted by Tsehayae and Fayek (2014) initially identified 169 factors that influence CLP. They collected 112 factors influencing CLP for concrete placing activities over 92 days of data collection. Based on Tsehayae and Fayek (2014), Table 3.1 shows the category, project types, activities studied, description of the activities, and the number of data instances collected. In this study, CLP is defined as a ratio of output, which is installed quantity, to input, which is labour work hours; CLP has positive real values.

Table 3.1. Studied activities for CLP modeling

Trade category	Project types	Activity	Activity description	Data instances
Concreting	Commercial mixed-use office-staff facility building, industrial warehouse building, commercial warehouse building, mixed residential-community center building, high-rise mixed commercial-residential building, institutional building	Columns	Concrete placement for columns	21
		Footings	Concrete placement for footings	5
		Grade beams	Concrete placement for grade beams	6
		Pile caps	Concrete placement for pile caps	2
		Slabs	Concrete placement for slabs	28
		Walls	Concrete placement for walls	30

In the existing data set, some CLP factors are objective, such as crew size, which has a numerical measure (in terms of number of workers), while other factors are subjective, such as complexity of task, which does not have a well-defined measurement. Subjective factors were measured using a predetermined rating scale of 1–5, according to Tsehayae and Fayek (2016b). CLP factors can be grouped into six levels: (1) activity, (2) project, (3) organizational, (4) provincial, (5) national, and (6) global.

3.2.2. CLP data preparation

As real data may be incomplete and noisy, data preparation is an essential preprocessing step for data mining. Data preparation produces a data set smaller than the original one, which can improve the efficiency of data analysis and prediction.

The CLP data preparation process consists of normalization, imputing missing values, removing factors with zero variance, and eliminating outliers.

3.2.2.1. Normalization

Mostly, CLP data have varying scales that lead to increased training time and biases in predictive models and affect convergence in prediction (Golnaraghi et al. 2020). Hence, the experimental data are normalized using Equation (3.1) in a process called "max–min normalization", where x_{ij} is the value of instance i from factor j; x_{jmin} and x_{jmax} are the minimum and maximum values of factor j, respectively; and r_{ij} is the normalized value of instance i from factor j. Max–min normalization guarantees that all features have the exact same scale.

$$r_{ij} = \frac{x_{ij} - x_{jmin}}{x_{jmax} - x_{jmin}} \tag{3.1}$$

3.2.2.2. Impute missing values

Data sets often have some missing values, due to human error or non-availability of real data. Imputation methods use ML algorithms to help estimate missing values. Based on Choudhury and Pal (2019), the neural network-based imputation method is able to train a data set containing incomplete samples and identify instances similar to instances with missing values. Based on the results of several studies (Choudhury and Pal 2019; Nelwamondo et al. 2013; Yuan et al. 2018), neural network imputation was applied in the present study in order to impute missing values of CLP.

3.2.2.3. Remove factors with zero deviation

Standard deviation is a measure of the variance of each factor in a data set. Removing factors with no variation in data instances is a pre-processing step for data sets (Xu et al. 2019). In this study, CLP factors with zero standard deviation were removed from the data set.

3.2.2.4. Eliminate outliers

Detecting and eliminating outliers is another essential step in data preparation. Although outliers are part of a data set, they are significantly different from other observations. In this study, Tukey's method, which utilizes the median, upper, and lower quartiles of a data set, was applied as an outlier detection method. Since quartiles are resistant to farthest data of the data set, Tukey's method is less sensitive, compared to methods using mean and standard variance (Sandbhor and Chaphalkar 2019).

3.2.3. HFS

The developed HFS is a combination of the ReliefF algorithm as a filter method and the integration of SVM and GA as a wrapper method and is utilized to identify the factors that are most predictive of CLP. The structure of three algorithms, namely ReliefF, SVM, and GA, are briefly discussed in the following sections.

3.2.3.1. ReliefF algorithm

Relief is one of the widely used filter based feature selection methods that identifies the best subset of features by measuring features' weights. This algorithm was proposed by Kira and Rendell (1992) which assigns weights to features based on the correlation between features and categories and also selects all features with greater weight than an artificial threshold. Notably, Relief algorithm is limited to binary classification problems. To address this problem, ReliefF algorithm was introduced by Kononeko (1994), which has the ability of working with multiclass problems. ReliefF is a distance based feature selector, which uses Manhattan distance to measure weights. The evaluation criteria of ReliefF algorithm is presented in Equation (3.2), where $W(f_{0,i})$ acts for the weight of ith feature before updating; $W(f_i)$ is the updated weight of ith feature; A is the vector of features; k is the number of nearest neighbors; ; m is the number of cycles, $f_{h(x_i)}$ and $f_{r(C)}$ are the value of k nearest neighbors of x_i in the same and different class, respectively, $P(C)$ is the ratio of the target samples C to the total sample; $P(class(x_i))$ is the ratio of the samples in the same class including x_i to the total samples; and $diff$ () denotes the distance of two samples on each feature in A.

$$W(f_i) = W(f_{0,i}) - \frac{\sum_{j=1}^{k} diff(A, x_i, f_{h(x_i)})}{m \times k} + \sum_{C \neq class(x_i)} \frac{P(C)}{1 - P(class(x_i))} \times \frac{\sum_{j=1}^{k} diff(A, x_i, f_{r(C)})}{m \times k} \quad (3.2)$$

This study uses the Manhattan distance to measure distance between two samples, as it is shown in Equation (3.3).

$$diff(A, R1, R2) = \frac{|R1 - R2|}{A_{max} - A_{min}} \tag{3.3}$$

In this study, ReliefF selected the most correlated CLP factors as its output. The factors selected by ReliefF were then applied as inputs to the combination of SVM and GA.

3.2.3.2. Support vector machine

SVMs can solve linear and non-linear problems and can provide power classification results (Mathur and Foody 2008). The most important advantage of SVM is that it can control the over learning and high dimensionality and decrease computational complexity and local extremum (Tao et al. 2019). For non-linear problems, by using mapping function, the data typically convert to a higher-dimensional dataset, which changes the problem to a linear and separable problem. By introducing Kernel function, the solving process of this kind of problems is facilitated. There are various types of kernel function namely, linear, polynomial and sigmoid functions. However, radial basis function (RBF), which is presented in Equation (3.4), is the most popular kernel function because it requires only one parameter, δ, which is a free parameter with a significant effect on classification accuracy and has a lower complexity in comparison with other functions. Another essential parameter in SVM problems is C, which is the penalty factor and shows the cost of misclassification. According to the significance of C and δ on the result of SVM, they needed to be optimized for obtaining the desired accuracy, which can be done by using GA optimization.

$$K(x, y) = \frac{\exp(-|x-y|^2)}{\delta^2} \tag{3.4}$$

3.2.3.3. Genetic Algorithm (GA)

GA is an adaptive heuristic search algorithm, which is looking for optimal solution as a goal. GA operates with population, and it is inspired by the mechanism of natural selection and natural genetics. GA uses a fitness function to estimate the significance of the result in the evaluation step. Two GA operators, mutation and crossover functions, randomly transfer chromosomes and affect the fitness value. Crossover, specifies two chromosomes those will generate a new offspring chromosome. However, mutation is the process used to change genes in chromosomes from their initial state. Chromosomes will go through a mutation operation after the crossover process and a new offspring is generated (Bean 1994). Elitism as another GA process, copying a small part of

the fittest candidates to the next generation. This study selects four best chromosomes to be part of the next generation and a single-point crossover and binary mutation were used.

The GA minimizes the value of the fitness function, which is shown as FF and calculated for each chromosome by using Equation (3.5). SVM_RMSE is the root mean square error (RMSE) of SVM model, w is a weight of the specified number of factors (n_f), s_i is '1' if the factor i is selected or '0' if the factor i is not selected, and c_i is the cost of factor i.

$$FF = SVM\_RMSE \times (1 + w \times (\sum_{i=1}^{n_f} c_i \times s_i)) \quad (3.5)$$

3.2.3.4. HFS process

An overview of the presented HFS method is shown in Figure 3.2, which presents the process of integrating ReliefF as a filter method, and GA and SVM as the wrapper method.

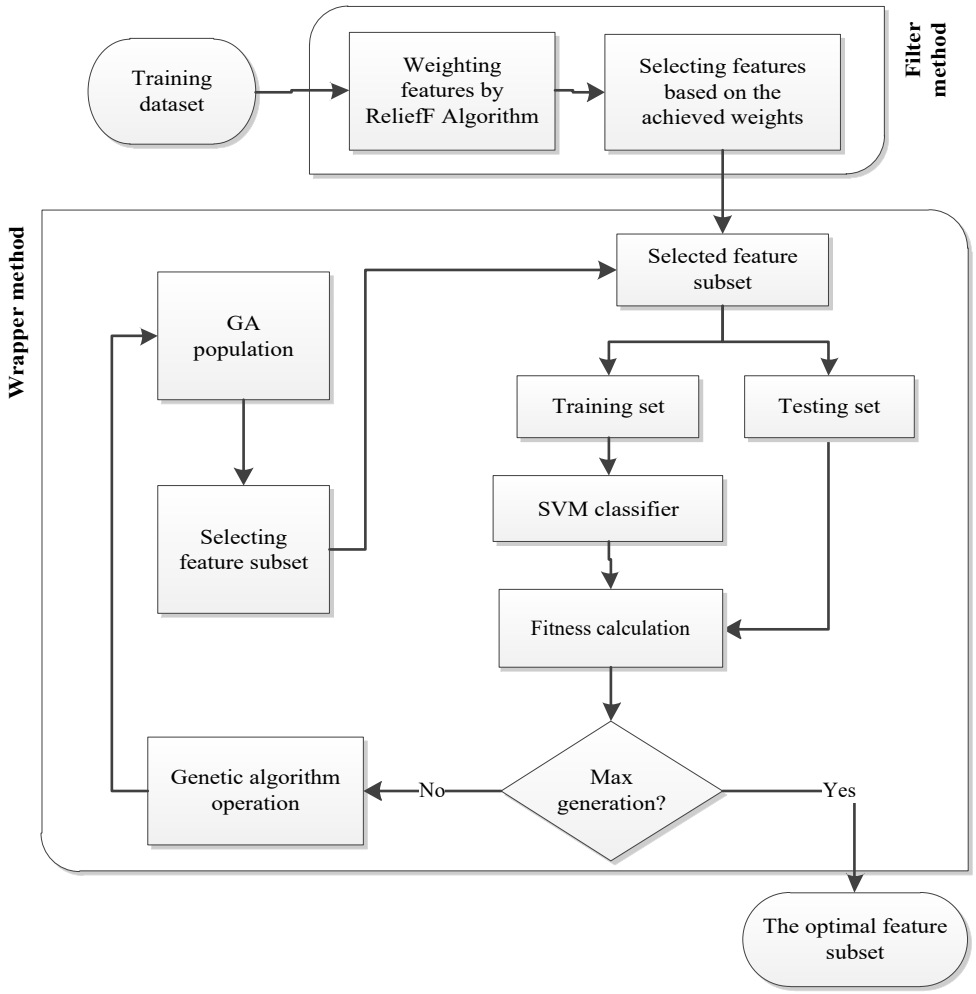

Figure 3.2. The overview of HFS method

As shown in Figure 3.2, the detailed explanation of the steps for developing the HFS method are as follows.

Step 1 – The ReliefF algorithm evaluates the weight of each feature according to the correlations between features and ranks them in terms of their weights. After the ReliefF process is complete, feature weights (w_r) are normalized from 0 to 1 to make the wrapper process more effective; by using a defined threshold (τ) in the range 0–1, any features with a weight $w_r \geq \tau$ are selected. Note that not all features with a weight greater than the threshold (τ) will necessarily be relevant, because it is possible for some irrelevant features to have an acceptable weight (Urbanowicz et al. 2018).

Step 2 – GA generates the random initial population of chromosomes. Each chromosome in the population represents an available solution to the feature subset selection problem.

Step 3 –Selected features that have weights greater than the threshold are the input of the SVM classifier.

Step 4 – The training set and testing set are built from the selected feature dataset. Then, using the training set, the process of training the SVM begins, while the testing set is utilized to calculate the SVM error.

Step 5 – The fitness function calculation process is completed using the calculated RMSE for SVM classification according to Equation (3.5).

Step 6 – If termination criteria are satisfied, the process ends; otherwise, the process goes to the next generation by GA.

Step 7 – GA searches for better solutions by using crossover, mutation, and elitism. In this study, single-point crossover and binary mutation were performed. Also, per the elitism process the four best chromosomes are selected to be part of the population in the next generation.

Once the final generation meets termination criteria, the iteration stops, and the selected feature subset is the one that has the best predictor of CLP among all feature subsets. The termination

criteria are: either the generation number reaches a determined value, or the fitness value does not improve during a specified number of generations.

3.2.3. CLP predictive modeling

According to the literature review on past CLP modeling techniques, ANN and ANFIS have been found to perform well and thus were chosen for this study. ANN is a suitable model for complex relationships between CLP and the factors that influence it, as these relationships cannot be obtained in a precise manner (El-Gohary et al. 2017; Song and AbouRizk 2008). ANFIS models have been widely used in past CLP studies because of their superiority in being less reliant on expert knowledge and having a systematic data-driven process (Sarihi et al. 2021). In order to optimize ANFIS parameters, the integration of ANFIS and GA was also developed. Another algorithm that shows accurate performance in a number of studies in other disciplines is RF, which was developed and compared with the other techniques in this study. Results from past studies show that RF is highly capable of solving non-linear classification problems compared to other ML models (Momade et al. 2020). As most crucial factors related to CLP do not follow a normal distribution, RF is a common ML technique in modeling construction productivity (Liu et al. 2018). The following sections discuss the structure and components of these four widely used ML modeling techniques and developed in this study.

3.2.3.1. Artificial neural network (ANN)

In the past few decades, ANN has become popular and helpful model for classification, clustering, pattern recognition, and prediction in many disciplines (Taheri et al. 2017). The potential of ANNs is the high-speed processing provided in a massive parallel implementation. ANNs have the ability to learn from experience to enhance their performance and adapt themselves to changes in the environment. ANNs are able to deal with noisy or incomplete data and can be very effective, mostly in problems where the relationships between inputs and outputs are not sufficiently known (Almási et al. 2016). So, based on their abilities ANNs can be ideal alternatives for modeling productivity. ANN consists of three layers: input layer, hidden layers, and output layer. Based on Boussabaine et al. (1996), ANNs mainly comprises the following components: (1) a set of neurons, (2) a connection pattern among the neurons, (3) each neuron's state of activation, (4) the activation rules, (5) the propagation method, and (6) a learning method. In this study, a multi-layer

feedforward back-propagation network with one hidden layer is developed as an ANN model to predict CLP.

3.2.3.2. Adaptive neuro fuzzy systems (ANFIS)

ANFIS is a hybrid FIS that integrates the linguistic interpretability and fuzzy reasoning of FIS and learning capability of ANN in order to map inputs to an output (Siraj et al. 2016). In an ANFIS structure, fuzzy rules are extracted from ANN and the parameters of fuzzy membership functions are adaptively utilized during the hybrid learning process (Moayedi et al. 2020).

3.2.3.3. ANFIS-GA

The combination of ANFIS and GA, is presented to improve the performance of the ANFIS model and optimize its parameters. GA is utilized to find the optimum parameters of ANFIS. Figure 3.3 shows the scheme of ANFIS-GA model. The population size, maximum iteration, mutation and crossover rate of the model has significant effect on the performance.

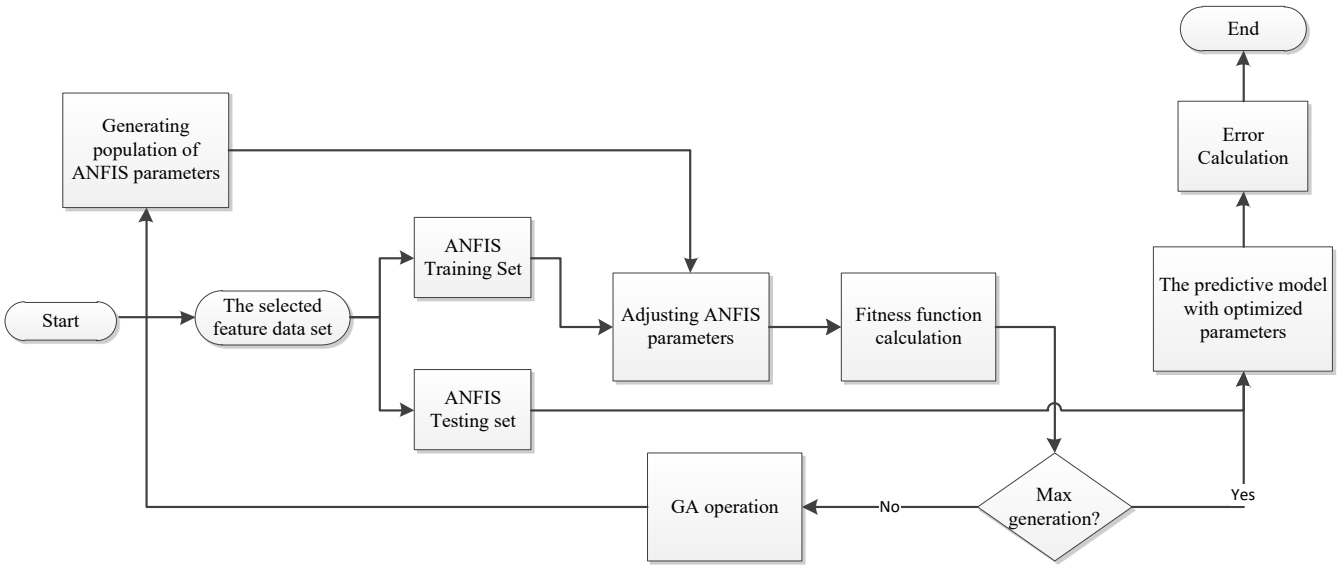

Figure 3.3. The ANFIS-GA structure

3.2.3.4. Random forest

RF could be considered an ensemble of classification and regression tree (CART), since multiple CART models are generated and used as base models (Liu et al. 2018; Momade et al. 2020). In this approach, the non-correlated trees (T_b) with low bias and variance are generated (Breiman

2001). Then, RF algorithm integrates all the regression trees using the bagging method (Schapire 2003). By reducing the variance, bagging method is able to improve prediction accuracy. CART is an unstable learner as one change in learning data can change the first splitting parameter, and consequently, change the tree structure. By using a set of trees instead of a single tree for prediction, RF prevails the instability of CART. RF tries to improve the diversity of trees using training data and input variables randomization. In this approach firstly, RF generates several training dataset by sampling randomly from the original training dataset. After generating new training datasets and before trees splitting process, RF implements variable randomization to boost the diversity of trees. Variable set randomization generates a random variable for each new training dataset. As both training data and variable sets are generated randomly, the trees in RF different from each other and also independent (Wang et al. 2018). Then, RF combines all trees by averaging their predictions. This joint prediction process increases the accuracy and decreases the large errors (Grandvalet 2004).

3.2.4. CLP optimization

In the last step of the methodology, the PSO algorithm searches for the optimum values of CLP and the factors influencing it, using the predictive model proposed in the previous section.

Swarm intelligence (SI) refers to a subset of artificial intelligence and it has been identified by Beni and Wang (1993) in the context of developing cellular robotic systems. There are several reasons responsible for the growing popularity of such SI-based algorithms, most importantly being the flexibility and versatility offered by these algorithms. Also, the self-learning capability and adaptability to external variations are the key features of these algorithms which has attracted immense interest and identified several application areas. PSO is one of the popular SI-based algorithms proposed by Kennedy and Eberhart (1995) for the first time. PSO is inspired by the behavior of flocks of birds, or swarm of insects in which individuals are called particles and the population is called a swarm. Although, PSO is simple to implement, it is able to find solutions with acceptable accuracy that makes it popular (Zheng et al. 2018; Sengupta et al. 2018). Each particle maintains three D-dimension vectors: position vector, velocity vector and personal best vector. Particles retain in their current position in position vector $X_i = (x_i^1, x_i^2, \dots, x_i^D)$, for $i = 1$ to N (N is the number of particles). Particles obtain their initial positions randomly in the search space. Velocity vector $V_i = (v_i^1, v_i^2, \dots, v_i^D)$ of ith particle is utilized to update its position and also

gets initial value randomly. The best position attained by ith particle is preserved in personal best vector is denoted as $Pbest_i = (Pbest_i^1, Pbest_i^2, \ldots, Pbest_i^D)$. Therefore, the swarm best position is considered as $Gbest = (Gbest^1, Gbest^2, \ldots, Gbest^D)$. The movement of the particle is related to updating its velocity and position attributes in the tth iteration (t = 2, 3 . . .), based on Equation 3.6 and Equation 3.7.

$$v_i^d(t+1) = wV_i^d(t) + c_1 r_1 \left(Pbest_i^d(t) - x_i^d(t)\right) + c_2 r_2 (Gbest_i^d(t) - x_i^d(t)) \quad (3.6)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (3.7)$$

where w is the inertia weight, c_1 is the cognitive acceleration coefficient, c_2 is the social acceleration coefficient, and r_1 and r_2 are random values between 0 and 1. Figure 3.4 presents a flowchart of PSO algorithm.

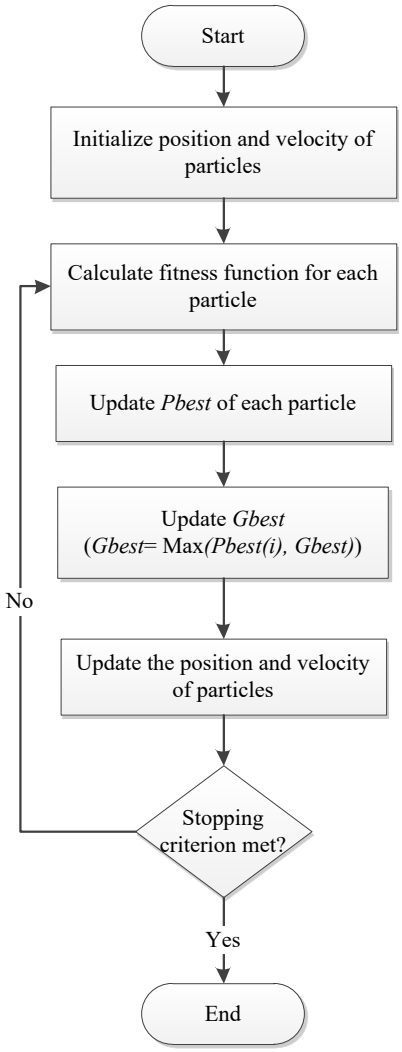

Figure 3.4. Overview of PSO

The objective of the optimization phase of this study contained the following goals:

- Goal 1: Predicted CLP (CLP_{Pred}) has minimum deviation from "targeted CLP" (CLP_{tgt}), as shown in Equation (3.8), where ω is the relative importance of Goal 1 compared to Goal 2.

$$\text{Goal 1} = \omega \times \left(CLP_{tgt} - CLP_{Pred}\right)^2 \tag{3.8}$$

- Goal 2: Predicted CLP factors (F_{Predi}) have minimum deviation from "average value of factors" (F_{Avgi}) in the dataset, among all the possible combinations of improvement scenarios, as shown in Equation (3.9).

$$\text{Goal 2} = (1 - \omega) \times \sum_{i=1}^{n} (F_{Predi} - F_{Avgi})^2 \tag{3.9}$$

In Goal 1, "targeted CLP" is the preferable CLP that a company tries to achieve. In this study, the value of CLP is between 0 and 1 after the normalization process, and greater CLP indicates better productivity in a project. Goal 1 tries to predict CLP considering the minimum distance from the targeted CLP.

Goal 2 tries to minimize changes in factors that most influence CLP. Companies mostly prefer minimum changes and corrective measures to achieve the preferable CLP because of the cost of implementing new strategies and corrective measures. In Goal 2, the average value of each factor is achieved from the existing CLP dataset, which is discussed in section 4. Since obtaining a value near the average value of each factor in the dataset is feasible, the goal is to have minimum distance between the average and optimum values for each factor. Therefore, the objective function is defined as in Equation (3.10):

$$Minimize \left(Z = \omega \times \left(CLP_{tgt} - CLP_{Pred}\right)^2 + (1 - \omega) \times \sum_{i=1}^{n} (F_{Predi} - F_{Avgi})^2 \right) \tag{3.10}$$

where CLP_{tgt} and CLP_{Pred} are the targeted CLP and predicted CLP, respectively, n is the number of selected factors affecting CLP, F_{Predi} is the predicted value of the ith CLP factor, F_{Avgi} is the average value of the ith CLP factor in the dataset, ω is the relative importance of Goal 1 compared to Goal 2, and Z is the minimum value of objective function. Objective function ranges from 0 to

1, where 0.5 means Goals 1 and 2 have equal importance. The outputs of this model will be CLP_{Pred}, which is the optimized and predicted CLP value, and F_{Predi}, which is the predicted value of factors influencing CLP.

3.3. References

Almási, A. D., Woźniak, S., Cristea, V., Leblebici, Y., and Engbersen, T. (2016). "Review of advances in neural networks: Neural design technology stack." *Neurocomputing*, 174, 31–41.

Atallah, D. M., Badawy, M., El-Sayed, A., and Ghoneim, M. A. (2019). "Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier." *Multimedia Tools and Applications*, 78(14), 20383–20407.

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

Beni, G., and Wang, J. (1993). "Swarm intelligence in cellular robotic systems." In *Robots and Biological Systems: Towards a New Bionics?*, Dario P., Sandini G., and Aebischer P., eds., 703–712. Berlin/Heidelberg, Germany: Springer-Verlag.

Boussabaine, A. H. (1996). "The use of artificial neural networks in construction management: A review." *Construction Management and Economics*, 14(5), 427–436.

Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5–32.

Cheng, M. Y., Cao, M. T., and Jaya Mendrofa, A. Y. (2021). "Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine." *Journal of Building Engineering*, 35, 101973.

Choudhury, S. J., and Pal, N. R. (2019). "Imputation of missing data with neural networks for classification." *Knowledge-Based Systems*, 182, 104838.

Ebrahimi, S., Raoufi, M., and Fayek, A. R. (2020). "Framework for integrating an artificial neural network and a genetic algorithm to develop a predictive model for construction labour productivity." *Construction Research Congress 2020*, American Society of Civil Engineers, Reston, VA, 58–66.

El-Gohary, K. M., Aziz, R. F., and Abdel-Khalek, H. A. (2017). "Engineering approach using ANN to improve and predict construction labour productivity under different influences." *Journal of Construction Engineering and Management*, 143(8), 04017045.

Golnaraghi, S., Moselhi, O., Alkass, S., and Zangenehmadar, Z. (2020). "Predicting construction

labour productivity using lower upper decomposition radial base function neural network." *Engineering Reports*, 2(2), 1–16.

Grandvalet, Y. (2004). "Bagging equalizes influence." *Machine Learning*, 55(3), 251–270.

Grau, D., Caldas, C. H., Haas, C. T., Goodrum, P. M., and Gong, J. (2009). "Assessing the impact of materials tracking technologies on construction craft productivity." *Automation in Construction*, 18(7), 903–911.

Kennedy, J., and Eberhart, R. (1995). "Particle swarm optimization." *Proceedings of ICNN'95 - International Conference on Neural Networks*, IEEE, 1942–1948.

Kira, K., and Rendell, L. A. (1992). "A practical approach to feature selection." *Machine Learning: Proceedings of the Ninth International Workshop (ML92) at the Ninth International Machine Learning Conference, Aberdeen, Scotland, 1992*, 249–256. Elsevier.

Kononenko, I. (1994). "Estimating attributes: Analysis and extensions of RELIEF." *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, Bergadano F., and De Raedt L. (eds.), volume 784, 171–182. Berlin/Heidelberg, Germany: Springer-Verlag.

Liu, X., Song, Y., Yi, W., Wang, X., and Zhu, J. (2018). "Comparing the random forest with the generalized additive model to evaluate the impacts of outdoor ambient environmental factors on scaffolding construction productivity." *Journal of Construction Engineering and Management*, 144(6), 04018037.

Mathur, A., and Foody, G. M. (2008). "Multiclass and binary SVM classification: Implications for training and classification users." *IEEE Geoscience and Remote Sensing Letters*, 5(2), 241–245.

Moayedi, H., Raftari, M., Sharifi, A., Jusoh, W. A. W., and Rashid, A. S. A. (2020). "Optimization of ANFIS with GA and PSO estimating α ratio in driven piles." *Engineering with Computers*, 36(1), 227–238.

Momade, M. H., Shahid, S., Hainin, M. R. bin, Nashwan, M. S., and Tahir Umar, A. (2020). "Modelling labour productivity using SVM and RF: a comparative study on classifiers performance." *International Journal of Construction Management*, 0(0), 1–11.

Nelwamondo, F. V., Golding, D., and Marwala, T. (2013). "A dynamic programming approach to missing data estimation using neural networks." *Information Sciences*, 49–58.

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." *Proceedings Asian Conference on Intelligent Information and Database Systems*, 721–727. Cham, Switzerland: Springer.

Sandbhor, S., and Chaphalkar, N. B. (2019). "Impact of outlier detection on neural networks based property value prediction." In Satapathy S., Bhateja V., Somanah R., Yang XS., Senkerik R., Eds., *Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing*, volume 862, 481–495. Singapore: Springer.

Sarihi, M., Shahhosseini, V., and Banki, M. T. (2021). "Development and comparative analysis of the fuzzy inference system-based construction labour productivity models." *International Journal of Construction Management*, 0(0), 1–18.

Schapire, R. E. (2003). "The boosting approach to machine learning: An overview." In Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), *Nonlinear Estimation and Classification* (pp. 149–171). New York, NY: Springer.

Sengupta, S., Basak, S., and Peters, R. A. (2018). "Particle swarm optimization: A survey of historical and recent developments with hybridization perspectives." *Machine Learning and Knowledge Extraction*, (1), 157–191.

Siraj, N. B., Fayek, A. R., and Tsehayae, A. A. (2016). "Development and optimization of artificial intelligence-based concrete compressive strength predictive models." *International Journal of Structural and Civil Engineering Research*, 5(3), 156–167.

Song, L., and Abourizk, S. M. (2008). "Measuring and modeling labour productivity using historical data." *Journal of Construction Engineering and Management*, 134(10), 786–794.

Taheri, K., Hasanipanah, M., Golzar, S. B., and Majid, M. Z. A. (2017). "A hybrid artificial bee colony algorithm-artificial neural network for forecasting the blast-produced ground vibration." *Engineering with Computers*, 33(3), 689–700.

Tao, Z., Huiling, L., Wenwen, W., and Xia, Y. (2019). "GA-SVM based feature selection and

parameter optimization in hospitalization expense modeling." *Applied Soft Computing*, 75, 323–332.

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., and Yildirim, M. B. (2018). "Predicting graft survival among kidney transplant recipients: A Bayesian decision support model." *Decision Support Systems*, 106, 97–109.

Tsehayae, A. A., and Fayek, A. R. (2014). "Identification and comparative analysis of key parameters influencing construction labour productivity in building and industrial projects." *Canadian Journal of Civil Engineering*, 41(10), 878–891.

Tsehayae, A. A., and Fayek, A. R. (2016a). "Developing and optimizing context-specific fuzzy inference system-based construction labour productivity models." *Journal of Construction Engineering and Management*, 142(7), 04016017.

Tsehayae, A. A., and Fayek, A. R. (2016b). "System model for analysing construction labour productivity." *Construction Innovation*, 16(2), 203–228.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). "Relief-based feature selection: Introduction and review." *Journal of Biomedical Informatics*, 85, 189–203.

Venkatesh, B., and Anuradha, J. (2019). "A hybrid feature selection approach for handling a high-dimensional data." *Proceedings of the 6th International Innovations in Computer Science and Engineering Conference*, edited by Saini H., Sayal R., Govardhan A., and Buyya R., pages 365–373. Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, volume 74. Springer: Singapore.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., and Ahrentzen, S. (2018). "Random forest based hourly building energy prediction." *Energy and Buildings*, 171, 11–25.

Xu, W., Jiang, L., and Yu, L. (2019). "An attribute value frequency-based instance weighting filter for naive Bayes." *Journal of Experimental and Theoretical Artificial Intelligence*, 31(2), 225–236.

Yuan, H., Xu, G., Yao, Z., Jia, J., and Zhang, Y. (2018). "Imputation of missing data intime series for air pollutants using long short-term memory recurrent neural networks." *Proceedings 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive*

and Ubiquitous Computing and Wearable Computers, 1293–1300.

Zheng, Z., Saxena, N., Mishra, K. K., and Sangaiah, A. K. (2018). "Guided dynamic particle swarm optimization for optimizing digital image watermarking in industry applications." *Future Generation Computer Systems*, 88, 92–106.

Chapter 4. Experimental results and discussion[3]

4.1. CLP Data Preparation and Feature Selection

Based on the data preparation process, the CLP value was normalized using Equation (3.1). As a result of normalization, the CLP value is between 0 and 1, and greater CLP indicates better labour productivity for the project. After imputing missing values and removing factors with zero standard deviation, the number of factors was reduced to 108. By eliminating outliers from the CLP dataset, 7 data points were removed as outliers, and the total number of data points became 85. Therefore, the CLP dataset after the preparation process had 85 data points, 108 CLP factors, and a CLP value.

Next, the number of features was reduced by the proposed HFS method. For this study, the threshold of 0.25 was defined for ReliefF. All features with weights greater than or equal to 0.25 were selected as essential features in the next HFS stage. From 108 factors in the final CLP dataset, ReliefF selected 43 as essential features. In the next stage of HFS, which is the integration of SVM and GA as a wrapper method, the GA parameter settings were a population size of 50, GA maximum iteration of 60, crossover rate of 0.83, and mutation rate of 0.2. SVM penalty factor C was 10, kernel type was RBF, and kernel cache was 200. These parameters were obtained by trial and error and are the optimum values for this case. The termination criteria were a maximum of 60 generations or no improvement of performance over 5 generations. The proposed wrapper method was developed considering these parameters, and it selected 14 of the 43 factors identified by ReliefF. After running the model multiple times, the set of 14 factors was selected when the RMSE of the run of the model was lowest. Table 4.1 presents the selected CLP factors resulting from HFS. As shown in Table 4.1, the first 11 factors are all from the activity level, and the next 3 factors belong to the project level, which shows the significant impact of activity-level factors on predicting CLP. From the selected factors, “Level of interruption and disruption,” “Complexity of task,” “Working condition (dust and fumes),” “Location of work scope (elevation),” and “Congestion of work area” are factors that negatively influence CLP. In other words, after normalization, when negatively influencing factors have values close to zero, they result in greater

[3] Parts of this chapter have been accepted for publication: Ebrahimi, S., Fayek, A. R., and Sumati, V. (2021). “Hybrid artificial intelligence HFS-RF-PSO model for construction labour productivity prediction and optimization.” *Algorithms*, 14(7), 214.

CLP, compared to when their values are close to 1. The other selected factors are positively influencing factors, and when their values are close to 1, they result in greater CLP.

Table 4.1. Input factors for CLP modeling

Selected factor	Scale of measure
(1) Crew size	Integer (Total number of crew members)
(2) Crew composition	Proportion (Ratio journeyman to apprentice to helper)
(3) Treatment of craftsperson by foreman	1–5 Predetermined rating
(4) Craftsperson trust in foreman	1–5 Predetermined rating
(5) Level of interruption and disruption	Integer (Number of interruptions and disruptions per day)
(6) Complexity of task	1–5 Predetermined rating
(7) Working condition (dust and fumes)	1–5 Predetermined rating
(8) Location of work scope (elevation)	Real number (elevation, m)
(9) Congestion of work area	Real number (ratio of actual peak manpower to actual average manpower)
(10) Fairness in performance review of crew by foreman	1–5 Predetermined rating
(11) Ground conditions	1–5 Predetermined rating
(12) Quality audits	Real number (Number of inspections per month)
(13) Risk monitoring and control	1–5 Predetermined rating
(14) Crisis management	1–5 Predetermined rating

4.2. CLP Modeling Comparison and Results

To develop the predictive CLP model, four different AI models were developed using the selected factors from HFS as input variables and CLP as the output. In order to avoid overfitting and manage the possible variations of input data, ten-fold cross validation is used for developing the classification models by partitioning the data into 10 random subsets (Huang et al. 2018). One subset is utilized to validate the model trained by the remaining subsets. This procedure is repeated 10 times such that each subset is used once for validation, and finally, the subset with minimum error is selected.

The accuracy of the four models was measured by comparing their predictions to the actual field data and calculating two commonly used error measures, mean absolute error (MAE) and RMSE, which are shown in Equation (4.1) and Equation (4.2), where t_i and y_i are the actual and predicted CLP values for the ith instance, respectively, and m is the number of instances. For this purpose, data were divided into training and testing datasets, in which 70% of data are used for training and 30% for testing.

$$RMSE = \sqrt{\sum_i (t_i - y_i)^2/m} \tag{4.1}$$

$$MAE = (\sum_i |y_i - t_i|)/m \tag{4.2}$$

For development of the ANN model, using MATLAB NN Toolbox, a multilayer feedforward back-propagation network with two hidden layers was considered, and the hidden layer sizes were 5 and 6. The learning rate was set to 0.33, and 200 training cycles were performed. The ANN model resulted in an RMSE of 0.164 and MAE of 0.130 for the training dataset, and an RMSE of 0.165 and MAE of 0.135 for the testing dataset.

The ANFIS model was generated using the ANFIS function of MATLAB Fuzzy Logic Toolbox. The basic learning rules for optimizing membership functions in ANFIS are either hybrid learning or back-propagation gradient descent. Hybrid learning combines the gradient descent and least square methods, and it overcomes the major limitation of the back-propagation method, which is that the learning process gets trapped in local minima. Therefore, this thesis used the hybrid

learning method. The training dataset was grouped using subtractive clustering with an influence range of 0.4, squash factor of 1.15, and accept and reject ratios set at 0.5 and 1.15, respectively. The selected CLP factors were used as input variables and CLP as the output of ANFIS. The ANFIS model resulted in an RMSE of 0.042 and MAE of 0.034 for the training dataset and an RMSE = 0.176 and MAE = 0.138 for the testing dataset.

The ANFIS-GA model, developed using MATLAB, tries to optimize ANFIS parameters, and it showed better performance than ANFIS alone. In this study, the values of 0.2, 0.83, and 60 were assigned for the mutation rate, crossover percentage, and maximum iteration of GA, respectively. These parameters are obtained by trial and error and are the optimum values for this case. Different sizes were tested to find the appropriate population size and based on the results as shown in Table 4.2 the ANFIS-GA model with a population size of 25 had the best testing performance, which included an MAE of 0.096 for the training dataset and MAE of 0.129 for the testing dataset. Therefore, a population size of 25 was used in this study.

Table 4.2. Selecting the population size in ANFIS-GA modeling

ANFIS-GA model no.	Population size	RMSE	
		Training	Testing
1	12	0.159	0.185
2	18	0.165	0.191
3	25	0.162	0.172
4	30	0.163	0.19

The RF model was developed using Python language programming and required three parameters, namely the minimum number of terminal nodes for each tree, the number of trees, and the number of randomly selected variables to grow the trees (Wang et al. 2018). In this study, these three parameters were set to 5, 145, and 6, respectively. The results of the RF prediction model are listed in Table 4.3 along with results of the ANN, ANFIS, and ANFIS-GA models for comparison.

The results presented in Table 4.3 indicate the RF model had the highest accuracy among the four predictive models, with an RMSE of 0.137 and MAE of 0.112 in the testing dataset. The second

most accurate algorithm was the ANN model, with a testing dataset RMSE of 0.165 and MAE of 0.135. The third most accurate algorithm was the combination of ANFIS and GA, with an RMSE of 0.172 and MAE of 0.129 in the testing dataset. Finally, testing dataset RMSE of 0.176 and MAE of 0.138 indicate the ANFIS model was the least accurate.

Table 4.3. Comparing the performance of the four developed models for predicting CLP

Model	Training Dataset		Testing Dataset	
	RMSE	MAE	RMSE	MAE
ANN	0.164	0.130	0.165	0.135
ANFIS	0.042	0.034	0.176	0.138
ANFIS-GA	0.162	0.096	0.172	0.129
RF	0.074	0.051	0.137	0.112

According to the RMSE value of 0.137 for the RF testing dataset, CLP predicted by RF was closer to the actual CLP values than for the other three developed models. In other words, RF was found to be better than ANN, ANFIS, and ANFIS-GA in mapping the relationship between the selected CLP factors and CLP. Moreover, the closeness of the RMSE values for the training and testing datasets indicate that ANN and RF were more stable than ANFIS and ANFIS-GA. Therefore, the RF model was selected to predict CLP in the optimization process for this study. Comparing the results of this study with past studies indicate that the RF predictive model has better performance. For example, Gerami Seresht et al. (Gerami Seresht et al. 2020) obtained an RMSE value of 0.22 for their proposed CLP predictive model, while in this study using the same dataset, the RMSE value of the RF model was 0.137. Therefore, the proposed CLP predictive model achieved better performance accuracy in CLP prediction compared with Gerami Seresht et al. (2020).

4.3. CLP Optimization

Next, the integration of RF and PSO was developed to achieve the optimum value of the selected factors and maximum CLP value, according to the objective function in Equation (3.10). For this case study, the average value of each factor (F_{Avgi}) and CLP after normalization are shown in Table 4.4, and the average CLP value for the dataset is 0.259.

Table 4.4. Average values of selected factors and CLP of the dataset

Selected factor and CLP	**Average value in normalized dataset (F_{Avgi})**
(1) Crew size	0.302
(2) Crew composition	0.289
(3) Treatment of craftsperson by foreman	0.569
(4) Craftsperson trust in foreman	0.518
(5) Level of interruption and disruption	0.162
(6) Complexity of task	0.500
(7) Working condition (dust and fumes)	0.218
(8) Location of work scope (elevation)	0.132
(9) Congestion of work area	0.438
(10) Fairness in performance review of crew by foreman	0.694
(11) Ground conditions	0.368
(12) Quality audits	0.832
(13) Risk monitoring and control	0.264
(14) Crisis management	0.634
CLP	**0.259**

4.3.3. Sensitivity analysis

For the purpose of illustrating a CLP improvement trend, a sensitivity analysis was carried out to show the influence of different values of input parameters (namely ω and CLP_{tgt}) on output

variables (CLP_{Pred} and Z) for understanding the impact of input parameters on model output. Table 4.5 shows the results of the sensitivity analysis, which indicates the value of Z and predicted CLP as outputs based on different values of ω and CLP_{tgt} as inputs of the RF-PSO model. The value of ω was changed between 0.27 and 1; ω =1 is the largest possible value for ω and indicates that Goal 2 has no impact on the model. The CLP_{tgt} is in the range of 0.45 to 1, and $CLP_{tgt} = 1$ is the largest possible value for CLP resulting from the normalization process. Figure 4.1 is based on the results in Table 4.5, which shows the value of CLP_{Pred} for different values of ω and CLP_{tgt}. For a specific CLP_{tgt}, by increasing ω, CLP_{Pred} increases, which shows the model sensitivity to ω, which is the relative importance of Goal 1. For $CLP_{tgt} = 0.45$ and CLP_{tgt} =0.6, the changes in CLP_{Pred} are much less given ω greater than 0.4. So, it can be concluded that when CLP_{tgt} is less than or equal to 0.6, the most appropriate value of ω is less than or equal to 0.4. This means the minimum deviation of F_{Predi} (predicted value of CLP factors) from F_{Avgi} (average value of CLP factors in the dataset) as a Goal 2 in Equation (10) has more weight compared to the minimum deviation of CLP_{Pred} from CLP_{tgt} as a Goal 1 in Equation (9).

Table 4.5. The results of sensitivity analysis

CLP(*tgt*) \ ω		0.27	0.4	0.5	0.6	0.73	1
0.45	Z	0.041	0.045	0.038	0.056	0.033	0.000
	CLP	0.374	0.43	0.441	0.439	0.448	0.449
0.6	Z	0.042	0.129	0.0496	0.078	0.055	0.000
	CLP	0.386	0.565	0.586	0.599	0.596	0.599
0.75	Z	0.057	0.049	0.079	0.124	0.116	0.001
	CLP	0.522	0.561	0.616	0.649	0.671	0.721
0.9	Z	0.071	0.19	0.184	0.186	0.157	0.032
	CLP	0.555	0.558	0.664	0.678	0.685	0.728
1	Z	0.152	0.146	0.205	0.189	0.162	0.054
	CLP	0.713	0.697	0.714	0.728	0.737	0.769

For selecting the most appropriate weight and targeted CLP, a company's preference is important. Most companies prefer minimum deviation from "average value of factors," which is feasible to reach, helps them decrease the number of corrective measures that are required, and thus reduces the cost of implementing corrective measures. Based on this, Goal 2 needs to have more weight

compared to Goal 1, which leads to selecting a value of ω less than or equal to 0.5 as a weight of Goal 1.

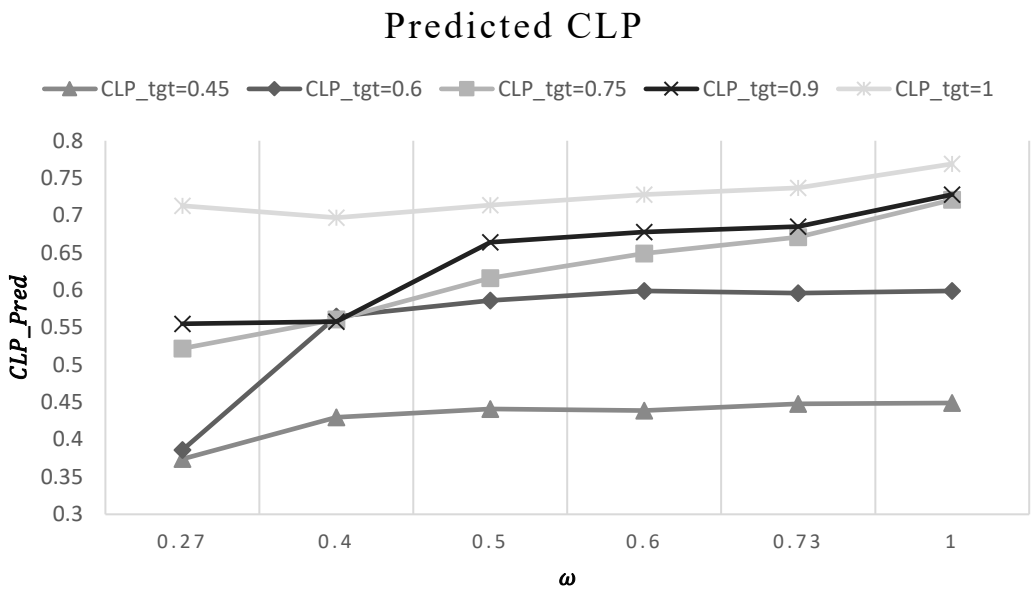

Figure 4.1. Predicted CLP from sensitivity analysis results

4.3.4. Optimization results

For this case study, the targeted CLP (CLP_{tgt}) of 0.75 and ω of 0.27 were selected. Equation (4.3) indicates the objective function of the HFS-RF-PSO algorithm according to the selected factors. In the presented algorithm, settings were number of particles = 50, maximum number of iterations = 30, and maximum velocity = 2, and the value of learning factors c_1 and c_2 were both set to 2.05. The initial values of the parameters were established on the basis of the relevant literature (El-Ghandour and Elbeltagi 2018). A large number of trials were performed to obtain the optimum values for this case.

$$Minimize \left(Z = 0.27 \times (0.75 - CLP_{Pred})^2 + 0.73 \times \sum_{i=1}^{14} (F_{Predi} - F_{Avgi})^2 \right) \quad (4.3)$$

Based on the selected inputs, the result of the RF-PSO model indicated 0.057 as a minimum value of Z, which is the minimum value of objective function (Equation (4.3)), and 0.522 was achieved as a maximum value of predicted CLP (CLP_{Pred}). The optimum value of each factor is shown in Table 4.6.

Table 4.6. Result of the RF-PSO algorithm for selected factors and CLP

Selected factor and CLP	**Optimum value (F_{Predi})**	**Deviation ($F_{Predi} - F_{Avgi}$)**
(1) Crew size	0.326	0.024
(2) Crew composition	0.364	0.075
(3) Treatment of craftsperson by foreman	0.587	0.018
(4) Craftsperson trust in foreman	0.535	0.017
(5) Level of interruption and disruption	0.043	-0.119
(6) Complexity of task	0.549	0.0490
(7) Working condition (dust and fumes)	0.108	-0.110
(8) Location of work scope (elevation)	0.176	0.044
(9) Congestion of work area	0.452	0.014
(10) Fairness in performance review of crew by foreman	0.808	0.114
(11) Ground conditions	0.372	0.004
(12) Quality audits	0.733	-0.099
(13) Risk monitoring and control	0.271	0.007
(14) Crisis management	0.629	-0.005
CLP	**0.522**	**0.263**

The optimum value of each factor was obtained from the RF-PSO model as the predicted values for CLP factors (F_{Predi}) and the deviation of the optimum value from the average value for each factor. In other words, deviation from average value was achieved using Equation (4.4):

$$Deviation = F_{Predi} - F_{Avgi} \quad (4.4)$$

As shown in Table 4.6, the optimum value of the factors "Ground condition," "Crisis management," and "Risk monitoring and control" have the least deviation from the average value of selected factors from dataset (F_{Avgi}) with values of 0.004, -0.005, and 0.007, respectively. Therefore, these factors do not need major changes to achieve the optimum CLP value, which is 0.522. It is notable in Table 4.6 that the optimum values of "Level of interruption and disruption," "Working condition (dust and fumes)," and "Fairness in performance review of crew by foreman" have the largest deviation from the average value of the factors, which are -0.119, -0.110, and 0.114, respectively. In other words, "Level of interruption and disruption" needs to be reduced to 0.043, "Working condition (dust and fumes)" needs to be reduced to 0.108, and "Fairness in performance review of crew by foreman" needs to increase to 0.808 in order to obtain the optimum CLP value. Improving factors with high deviation helps companies reach optimum predicted CLP. In order to improve factors that have a high deviation from their average value, a number of improvement strategies and corrective measures can be implemented. For example, for reducing dust and fumes in working area, preventive maintenance for air-conditioning system can be conducted.

The proposed HFS-RF-PSO model has the potential to benefit construction companies in achieving their preferred labour productivity by applying the minimum changes to factors influencing CLP. Another capability of the proposed model is that companies can define their targeted value for each factor influencing CLP instead of the average value of factors. The results of the model will give them the values of predicted CLP and predicted factors in regard to having the minimum deviation from their targeted values for CLP as well as each factor. This novel approach can help companies identify factors that need the most changes for achieving their targeted CLP and, consequently, to prioritize the management practices that focus on factors with the greatest deviation from average value in the HFS-RF-PSO model.

4.4. References

El-Ghandour, H. A., and Elbeltagi, E. (2018). "Comparison of five evolutionary algorithms for optimization of water distribution networks." *Journal of Computing in Civil Engineering*, 32(1), 04017066.

Gerami Seresht, N., Lourenzutti, R., and Fayek, A. R. (2020). "A fuzzy clustering algorithm for developing predictive models in construction applications." *Applied Soft Computing*, 96, 106679.

Huang, Z., Yang, C., Zhou, X., and Huang, T. (2018). "A Hybrid Feature Selection Method Based on Binary State Transition Algorithm and ReliefF." IEEE Journal of Biomedical and Health Informatics, 23(5), 1888–1898.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., and Ahrentzen, S. (2018). "Random forest based hourly building energy prediction." *Energy and Buildings*, 171, 11–25.

Chapter 5. Conclusions and Recommendations

5.1. Introduction

This chapter provides the research summary, academic contributions, and industrial contributions of this research. This chapter also discusses the limitations of the research and provides recommendations for future research and development.

5.2. Research Summary

This research aimed to fill the gaps in construction research on CLP prediction and optimization. An extensive review of past research in developing CLP predictive and optimization models revealed several gaps. **The first gap** is the fact that the construction literature on identifying factors most influential on CLP mostly relied on expert knowledge, which can be very changeable from time to time. Furthermore, the high-dimensional feature space of labour productivity often imposes a high computational cost as well as the risk of "overfitting" when classification is performed. Therefore, reducing the dimensionality of labour productivity data and finding the most influencing parameters for CLP is necessary, and this can be done by applying feature selection methods. **The second gap** is related to the fact that most past CLP modeling studies used filter methods for selecting the most influential factors (Bai et al. 2019; Gerami Seresht and Fayek 2018; Tsehayae and Fayek 2016). However, using wrapper or HFS is more appropriate for predictive modeling using AI techniques, because of their superior performance (Piao and Ryu 2017). **The third gap** is related to the lack of comparative analysis on developing CLP predictive models in order to identify the appropriate AI models that achieve an optimal prediction evaluation index. Despite the wide application of the predictive model of CLP for project planning and control, a predictive model in its sole application cannot offer construction companies the optimal combination of influencing factors for improving CLP. As a **fourth gap** in labour productivity modelling, no studies have considered a predictive model for finding the maximum value of CLP considering changes in the most influential factors by using optimization methods.

To fill the mentioned research gaps, the objectives of this research were achieved in four stages, as discussed below.

5.2.1. The first stage: Literature review

An extensive literature review was conducted on relevant topics as listed in Chapter 2. First, previous research on developing different feature selection methods, namely filter, wrapper, and HFS methods, were reviewed. Thereafter, previous research on identifying CLP factors was reviewed. Next, previous predictive modeling techniques for CLP prediction were reviewed.

5.2.2. The second stage: Feature selection modeling

The input parameters' large feature space consisting of the factors influencing CLP had to be reduced to maintain interpretability and accuracy of CLP prediction. As discussed in Chapter 3, data preparation steps including normalization, imputing missing values, removing factors with zero deviation, and eliminating outliers were applied to improve the efficiency of data analysis and prediction. Max-min normalization was carried out to normalize the CLP dataset, and then neural network-based imputation method was applied to impute missing values. Tukey's method was used to detect and eliminate outliers. Next, the integration of ReliefF algorithm as a filter method with SVM-GA as a wrapper method was presented as an HFS model for identifying the factors most influential on CLP.

5.2.3. The third stage: CLP predictive modeling

Different predictive models of CLP, namely ANN, ANFIS, ANFIS-GA, and RF, were developed to carry out a comparative analysis of CLP prediction. Based on the results as RMSE and MAE, CLP predicted by RF was closer to the actual CLP values than that for other three developed models. Therefore, the RF model was selected to predict CLP in the optimization process. As a validation process, the achieved RMSE value was compared with a previous study (Gerami Seresht et al. 2020) that developed a CLP predictive model using the same dataset.

5.2.4. The fourth stage: CLP optimization

As a last stage of the proposed methodology, the integration of RF and PSO was developed in Python® to achieve the maximum CLP value, considering the minimum deviation from a company's targeted CLP value and finding the optimum value of the selected factors based on minimizing their deviation from their average value in the dataset. Then, a sensitivity analysis was carried out to illustrate a CLP improvement trend.

5.3. Research Contributions

5.3.1. Academic contributions

The academic contributions of this research are:

- Development of an HFS model that contains the integration of ReliefF and SVM-GA for identifying the most predictive factors for CLP. The proposed HFS model is expected to enhance the accuracy of CLP prediction and identify the most predictive factors for CLP.
- Development of four different predictive models for CLP using ANN, ANFIS, ANFIS-GA, and RF and identifying the most accurate model based on a comparative analysis. The comparative analysis of four predictive models showed the RF model obtained better accuracy compared with the three other models.
- Development of a novel approach – the HFS-RF-PSO algorithm – for optimizing factors that influence CLP and identifying the maximum CLP value, considering the minimum deviation from targeted CLP value and finding the optimum value of the selected factors based on minimizing their deviation from their average value in the dataset. The proposed model can determine the maximum value of CLP and optimum value of each influential factor using optimization techniques.

5.3.2. Industrial Contributions

The industrial contributions of this research are:

- Identification of the most value-adding CLP factors, which helps construction planners identify strategies for improving the most value-adding factors. This finding provides construction practitioners with information about the factors that have the highest level of influence on predicting CLP.
- Prediction of labour productivity for use in construction project cost estimation and scheduling. The developed model can be used to provide reliable prediction of CLP values for concrete-placing activities.
- Development of a hybrid HFS-RF-PSO model for optimizing CLP and its factors, by considering a company's targeted CLP value, which helps project managers predict, optimize, and improve the CLP value by taking into account factors that are most predictive of CLP. Furthermore, the model will be effective for construction planners to carry out productivity improvement studies and analyze different scenarios. This approach can help

companies identify factors that need the most changes for achieving their targeted CLP and, consequently, to prioritize management practices that focus on factors with the greatest deviation from average value in the HFS-RF-PSO model.

- Facilitating the adoption of best practices. Although a number of best practices have been presented in past studies, substantiating the possible gain in CLP due to the adoption of such best practices remains difficult. This research developed a CLP hybrid model that can quantify expected gains in CLP due to the adoption of best practices, such as implementing labour productivity measurement practices or safety training, and the predicted gains in CLP can be further examined using case study projects.
- Although construction projects are unique and the factors affecting CLP may differ from project to project, the proposed model is flexible and generic, and new influencing factors can be added to the existing model structure to predict and optimize CLP and its factors for a given project.

5.4. Research Limitations and Recommendations for Future Research

The following limitations were encountered in the research study, and recommendations are suggested for future work.

1. The hybrid model was developed using field data collected for concrete-placing activities in the past study. However, in order to develop a generic model of CLP for different types of labour-dependent activities, new data need to be collected. Additional investigation of other labour-intensive activities, such as welding, piping, and scaffolding, is recommended to further improve the developed hybrid model. Further, another limitation of modeling CLP is that data collection still relies on experts since the large number of factors are subjective. Future data collection also needs to investigate scale of measure of each factor and define factors to be less reliant on other factors.
2. Although PSO algorithm is computationally efficient compared to other optimization techniques and is robust with respect to control parameters, it can fall into a local optimum in high-dimensional space. In future research, an adaptive PSO algorithm can be developed and added to the hybrid model to improve diversity of algorithm and avoid falling into local optimum.

3. The proposed hybrid model that has been developed for labour-intensive activities cannot accurately predict the productivity of equipment-intensive activities. Therefore, future research can focus on using the proposed methodology to model and optimize multifactor construction productivity, which includes labour, equipment, and materials.
4. This research has shown which factors need major improvement in order to achieve a CLP value close to the targeted CLP value. However, it does not present corrective measures relevant to the specified factors. Therefore, future studies can present corrective measures to improve CLP according to HFS-RF-PSO results that show which factors need the most changes for reaching targeted CLP. Future studies with case studies can further validate the proposed model for predicting CLP.
5. Many of the factors influencing CLP are subjective factors. As fuzzy models have the capability to deal with several subjective factors, and random forest shows the best performance compared to the other three models in this study, future studies can develop a fuzzy random forest model to deal with subjectivity of data as well.

5.5. References

Bai, S., Li, M., Kong, R., Han, S., Li, H., and Qin, L. (2019). "Data mining approach to construction productivity prediction for cutter suction dredgers." *Automation in Construction*, 105, 102833.

Gerami Seresht, N., and Fayek, A. R. (2018). "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management* (ASCE), 144(9), 04018091.

Gerami Seresht, N., Lourenzutti, R., and Fayek, A. R. (2020). "A fuzzy clustering algorithm for developing predictive models in construction applications." *Applied Soft Computing*, 96, 106679.

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." *Proceedings Asian Conference on Intelligent Information and Database Systems*, , 721–727. Cham, Switzerland: Springer.

Tsehayae, A. A., and Fayek, A. R. (2016). "Developing and optimizing context-specific fuzzy inference system-based construction labour productivity models." *Journal of Construction Engineering and Management*, 142(7), 04016017.

Bibliography

Agrawal, A., and Halder, S. (2020). "Identifying factors affecting construction labour productivity in India and measures to improve productivity." *Asian Journal of Civil Engineering*, 21(4), 569–579.

Ahmad, S. S. S., and Pedrycz, W. (2012). "Data and Feature Reduction in Fuzzy Modeling through Particle Swarm Optimization." *Applied Computational Intelligence and Soft Computing*, 2012, 1–21.

Alaghbari, W., Al-Sakkaf, A. A., and Sultan, B. (2019). "Factors affecting construction labour productivity in Yemen." *International Journal of Construction Management*, 19(1), 79–91.

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

Almási, A. D., Woźniak, S., Cristea, V., Leblebici, Y., and Engbersen, T. (2016). "Review of advances in neural networks: Neural design technology stack." *Neurocomputing*, 174, 31–41.

Atallah, D. M., Badawy, M., El-Sayed, A., and Ghoneim, M. A. (2019). "Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier." *Multimedia Tools and Applications*, 78(14), 20383–20407.

Bai, S., Li, M., Kong, R., Han, S., Li, H., and Qin, L. (2019). "Data mining approach to construction productivity prediction for cutter suction dredgers." *Automation in Construction*, 105, 102833.

Bean, J. C. (1994). "Genetic Algorithms and Random Keys for Sequencing and Optimization." *ORSA Journal on Computing*, 6(2), 154–160.

Beni, G., and Wang, J. (1993). "Swarm Intelligence in Cellular Robotic Systems." In *Robots and Biological Systems: Towards a New Bionics?*, Dario P., Sandini G., and Aebischer P., eds. , 703–712. Berlin/Heidelberg, Germany: Springer.

Boussabaine, A. H. (1996). "The use of artificial neural networks in construction management: A review." *Construction Management and Economics*, 14(5), 427–436.

Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5–32.

Canada, S. (2019). "Gross domestic product at basic prices, by industry, 2015–2019." <http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/manuf10-eng.html>.

Cheng, M.-Y., Cao, M.-T., and Jaya Mendrofa, A. Y. (2020). "Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine." *Journal of Building Engineering*, 101973.

Cheng, M. Y., Cao, M. T., and Jaya Mendrofa, A. Y. (2021). "Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine." *Journal of Building Engineering*, 35, 101973.

Chigara, B., and Moyo, T. (2014). "Factors affecting labor productivity on building projects in Zimbabwe." *International Journal of Architecture, Engineering and Construction*, 3(1), 57–65.

Choudhury, S. J., and Pal, N. R. (2019). "Imputation of missing data with neural networks for classification." *Knowledge-Based Systems*, 182, 104838.

Dash, M., and Liu, H. (1997). "Feature selection for classification." *Intelligent Data Analysis*, 1(3), 131–156.

Dehghan, R., Hazini, K., and Ruwanpura, J. (2015). "Optimization of overlapping activities in the design phase of construction projects." *Automation in Construction*, 59, 81–95.

Dixit, S., Mandal, S. N., Thanikal, J. V, and Saurabh, K. (2018). "Construction productivity and construction project performance in Indian construction projects." *Proceedings Creative Construction Conference 2018*, 379–386. Diamond Congress Ltd., Budapest University of Technology and Economics.

Doloi, H. (2008). "Application of AHP in improving construction productivity from a management perspective." *Construction Management and Economics*, 26(8), 841–854.

Durdyev, S., Ismail, S., and Kandymov, N. (2018). "Structural equation model of the factors affecting construction labor productivity." *Journal of Construction Engineering and Management*, 144(4), 04018007.

Eastman, C. M., and Sacks, R. (2008). "Relative Productivity in the AEC Industries in the United States for On-Site and Off-Site Activities." *Journal of Construction Engineering and Management* (ASCE), 134(7), 517–526.

Ebrahimi, S., Raoufi, M., and Fayek, A. R. (2020). "Framework for integrating an artificial neural network and a genetic algorithm to develop a predictive model for construction labor productivity." *Construction Research Congress 2020*, American Society of Civil Engineers, Reston, VA, 58–66.

El-Ghandour, H. A., and Elbeltagi, E. (2018). "Comparison of five evolutionary algorithms for optimization of water distribution networks." *Journal of Computing in Civil Engineering*, 32(1), 04017066.

El-Gohary, K. M., Aziz, R. F., and Abdel-Khalek, H. A. (2017). "Engineering approach using ANN to improve and predict construction labor productivity under different influences." *Journal of Construction Engineering and Management*, 143(8), 04017045.

Fei, Y., and Min, H. (2016). "Simultaneous feature with support vector selection and parameters optimization using GA-based SVM solve the binary classification." *2016 1st IEEE International Conference on Computer Communication and the Internet, ICCCI 2016*, 426–433.

Gerami Seresht, N., and Fayek, A. R. (2018). "Dynamic modeling of multifactor construction productivity for equipment-intensive activities." *Journal of Construction Engineering and Management* (ASCE), 144(9), 04018091.

Gerami Seresht, N., Lourenzutti, R., and Fayek, A. R. (2020). "A fuzzy clustering algorithm for developing predictive models in construction applications." *Applied Soft Computing*, 96, 106679.

Ghosh, M., Guha, R., Sarkar, R., and Abraham, A. (2019). "A wrapper-filter feature selection technique based on ant colony optimization." *Neural Computing and Applications*, 32, 7839–7857.

Golnaraghi, S., Moselhi, O., Alkass, S., and Zangenehmadar, Z. (2020). "Predicting construction labor productivity using lower upper decomposition radial base function neural network." *Engineering Reports*, 2(2), 1–16.

Grandvalet, Y. (2004). "Bagging equalizes influence." *Machine Learning*, 55(3), 251–270.

Grau, D., Caldas, C. H., Haas, C. T., Goodrum, P. M., and Gong, J. (2009). "Assessing the impact of materials tracking technologies on construction craft productivity." *Automation in Construction*, 18(7), 903–911.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature extraction: foundations and applications*. Berlin/Heidelberg, Germany: Springer-Verlag.

Hafez, S. M. (2014). "Critical factors affecting construction labor productivity in Egypt." *American Journal of Civil Engineering*, 2(2), 35.

Hai, D. T., and Van Tam, N. (2020). "Application of the regression model for evaluating factors affecting construction workers' labor productivity in Vietnam." *The Open Construction and Building Technology Journal*, 13(1), 353–362.

Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation). University of Waikato, Hamilton, New Zealand.

Heravi, G., and Eslamdoost, E. (2015). "Applying artificial neural networks for measuring and predicting construction-labor productivity." *Journal of Construction Engineering and Management*, 141(10), 04015032.

Hsu, H. H., Hsieh, C. W., and Lu, M. Da. (2011). "Hybrid feature selection by combining filters and wrappers." *Expert Systems with Applications*, 38(7), 8144–8150.

Huang, Z., Yang, C., Zhou, X., and Huang, T. (2018). "A Hybrid Feature Selection Method Based on Binary State Transition Algorithm and ReliefF." IEEE Journal of Biomedical and Health Informatics, 23(5), 1888–1898.

Irfan, M., Zahoor, H., Abbas, M., and Ali, Y. (2020). "Determinants of labor productivity for building projects in Pakistan." *Journal of Construction Engineering, Management & Innovation*, 3(2), 85–100.

Jarkas, A. M. (2015). "Factors influencing labour productivity in Bahrain's construction industry." *International Journal of Construction Management*, 15(1), 94–108.

Kari, T., Gao, W., Zhao, D., Abiderexiti, K., Mo, W., Wang, Y., and Luan, L. (2018). "Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm." *IET Generation, Transmission and Distribution*, 12(21), 5672–5680.

Kennedy, J., and Eberhart, R. (1995). "Particle swarm optimization." *Proceedings of ICNN'95 - International Conference on Neural Networks*, IEEE, 1942–1948.

Khanzadi, M., Nasirzadeh, F., Mir, M., and Nojedehi, P. (2017). "Prediction and improvement of labor productivity using hybrid system dynamics and agent-based modeling approach." *Construction Innovation*, 18(1), 2–19.

Kira, K., and Rendell, L. A. (1992). "A practical approach to feature selection." In *Machine Learning: Proceedings of the Ninth International Workshop (ML92) at the Ninth International Machine Learning Conference, Aberdeen, Scotland, 1992*, 249–256. Elsevier.

Kisi, K. P., Mani, N., Rojas, E. M., and Foster, E. T. (2017). "Optimal productivity in labor-intensive construction operations: Pilot study." *Journal of Construction Engineering and Management*, 143(3), 04016107.

Kononenko, I. (1994). "Estimating attributes: Analysis and extensions of RELIEF." In *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, Bergadano F., and De Raedt L. (eds.), volume 784, 171–182. Berlin/Heidelberg, Germany: Springer-Verlag.

Lee, C. P., and Leu, Y. (2011). "A novel hybrid feature selection method for microarray data analysis." *Applied Soft Computing*, 11(1), 208–213.

Lee, J., Park, Y. J., Choi, C. H., and Han, C. H. (2017). "BIM-assisted labor productivity measurement method for structural formwork." *Automation in Construction*, 84, 121–132.

Lin, C. L., and Lai, Y. C. (2020). "An improved time-cost trade-off model with optimal labor productivity." *Journal of Civil Engineering and Management*, 26(2), 113–130.

Liu, H., Zhou, M., and Liu, Q. (2019). "An embedded feature selection method for imbalanced data classification." *IEEE/CAA Journal of Automatica Sinica*, 6(3), 703–715.

Liu, X., Song, Y., Yi, W., Wang, X., and Zhu, J. (2018). "Comparing the random forest with the generalized additive model to evaluate the impacts of outdoor ambient environmental factors on scaffolding construction productivity." *Journal of Construction Engineering and Management*, 144(6), 04018037.

Loosemore, M. (2014). "Improving construction productivity: a subcontractor's perspective." *Engineering, Construction and Architectural Management*, 21(3), 245–260.

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., and Gao, Z. (2017). "A hybrid feature selection algorithm for gene expression data classification." *Neurocomputing*, 256, 56–62.

Lu, M. (2000). *Productivity studies using advanced ANN models* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses (204).

Mafarja, M., and Mirjalili, S. (2018). "Whale optimization approaches for wrapper feature selection." *Applied Soft Computing*, 62, 441–453.

Mathur, A., and Foody, G. M. (2008). "Multiclass and binary SVM classification: Implications for training and classification users." *IEEE Geoscience and Remote Sensing Letters*, 5(2), 241–245.

Mirahadi, F., and Zayed, T. (2016). "Simulation-based construction productivity forecast using neural-network-driven fuzzy reasoning." *Automation in Construction*, 65, 102–115.

Moayedi, H., Raftari, M., Sharifi, A., Jusoh, W. A. W., and Rashid, A. S. A. (2020). "Optimization of ANFIS with GA and PSO estimating α ratio in driven piles." *Engineering with Computers*, 36(1), 227–238.

Momade, M. H., Shahid, S., Hainin, M. R. bin, Nashwan, M. S., and Tahir Umar, A. (2020). "Modelling labour productivity using SVM and RF: a comparative study on classifiers performance." *International Journal of Construction Management*, 0(0), 1–11.

Monirul Kabir, M., Monirul Islam, M., and Murase, K. (2010). "A new wrapper feature selection approach using neural network." *Neurocomputing*, 73(16–18), 3273–3283.

Montaser, N. M., Mahdi, I. M., Mahdi, H. A., and Rashid, I. A. (2018). "Factors affecting construction labor productivity for construction of pre-stressed concrete bridges." *International Journal of Construction Engineering and Management*, 7(6), 193–206.

Nasirzadeh, F., Kabir, H. M. D., Akbari, M., Khosravi, A., Nahavandi, S., and Carmichael, D. G. (2020). "ANN-based prediction intervals to forecast labour productivity." *Engineering, Construction and Architectural Management*, 27(9), 2335–2351.

Nelwamondo, F. V., Golding, D., and Marwala, T. (2013). "A dynamic programming approach to missing data estimation using neural networks." *Information Sciences*, 49–58.

Nguyen, B. H., Xue, B., and Zhang, M. (2020). "A survey on swarm intelligence approaches to

feature selection in data mining." *Swarm and Evolutionary Computation*, 54(October 2019), 100663.

Parthasarathy, M. K., Murugasan, R., and Vasan, R. (2018). "Modelling manpower and equipment productivity in tall residential building projects in developing countries." *Journal of the South African Institution of Civil Engineering*, 60(2), 23–33.

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." *Proceedings Asian Conference on Intelligent Information and Database Systems*, 721–727. Cham, Switzerland: Springer.

Raoufi, M., and Fayek, A.R. (2018). "Fuzzy agent-based modeling of construction crew motivation and performance." *Journal of Computing in Civil Engineering*, 32(5), 04018035.

Sandbhor, S., and Chaphalkar, N. B. (2019). "Impact of outlier detection on neural networks based property value prediction." In Satapathy S., Bhateja V., Somanah R., Yang XS., Senkerik R., Eds., *Information Systems Design and Intelligent Applications*. *Advances in Intelligent Systems and Computing*, volume 862, 481–495. Singapore: Springer.

Sarihi, M., Shahhosseini, V., and Banki, M. T. (2021). "Development and comparative analysis of the fuzzy inference system-based construction labor productivity models." *International Journal of Construction Management*, 0(0), 1–18.

Schapire, R. E. (2003). "The boosting approach to machine learning: An overview." In Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), *Nonlinear Estimation and Classification* (pp. 149–171). New York, NY: Springer.

Sengupta, S., Basak, S., and Peters, R. A. (2018). "Particle swarm optimization: A survey of historical and recent developments with hybridization perspectives." *Machine Learning and Knowledge Extraction*, (1), 157–191.

Siraj, N. B., Fayek, A. R., and Tsehayae, A. A. (2016). "Development and optimization of artificial intelligence-based concrete compressive strength predictive models." *International Journal of Structural and Civil Engineering Research*, 5(3), 156–167.

Song, L., and Abourizk, S. M. (2008). "Measuring and modeling labor productivity using historical data." *Journal of Construction Engineering and Management*, 134(10), 786–794.

Taheri, K., Hasanipanah, M., Golzar, S. B., and Majid, M. Z. A. (2017). "A hybrid artificial bee colony algorithm-artificial neural network for forecasting the blast-produced ground vibration." *Engineering with Computers*, 33(3), 689–700.

Talhouni, B. (1990).*Measurement and analysis of construction labour productivity* (Doctoral dissertation). University of Dundee, Dundee, Scotland.

Tao, Z., Huiling, L., Wenwen, W., and Xia, Y. (2019). "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling." *Applied Soft Computing*, 75, 323–332.

Thomas, A. V., and Sudhakumar, J. (2014). "Modelling masonry labour productivity using multiple regression." *Proceedings 30th Annual Association of Researchers in Construction Management Conference, ARCOM 2014*, 1345–1354.

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., and Yildirim, M. B. (2018). "Predicting graft survival among kidney transplant recipients: A Bayesian decision support model." *Decision Support Systems*, 106, 97–109.

Tsehayae, A. (2015). *Developing and optimizing context-specific and universal construction labour productivity models* (Doctoral dissertation). University of Alberta, Edmonton, Alberta.

Tsehayae, A. A., and Fayek, A. R. (2014). "Identification and comparative analysis of key parameters influencing construction labour productivity in building and industrial projects." *Canadian Journal of Civil Engineering*, 41(10), 878–891.

Tsehayae, A. A., and Fayek, A. R. (2016a). "Developing and optimizing context-specific fuzzy inference system-based construction labor productivity models." *Journal of Construction Engineering and Management*, 142(7), 04016017.

Tsehayae, A. A., and Fayek, A. R. (2016b). "System model for analysing construction labour productivity." *Construction Innovation*, 16(2), 203–228.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). "Relief-based feature selection: Introduction and review." *Journal of Biomedical Informatics*, 85, 189–203.

Venkatesh, B., and Anuradha, J. (2019). "A hybrid feature selection approach for handling a high-dimensional data." *Proceedings of the 6th International Innovations in Computer Science and Engineering Conference*, edited by Saini H., Sayal R., Govardhan A., and Buyya R., pages 365–373. Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, volume 74. Springer: Singapore.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., and Ahrentzen, S. (2018). "Random forest based hourly building energy prediction." *Energy and Buildings*, 171, 11–25.

Wei, G., Zhao, J., Feng, Y., He, A., and Yu, J. (2020). "A novel hybrid feature selection method based on dynamic feature importance." *Applied Soft Computing*, 93, 106337.

Xu, W., Jiang, L., and Yu, L. (2019). "An attribute value frequency-based instance weighting filter for naive Bayes." *Journal of Experimental and Theoretical Artificial Intelligence*, 31(2), 225–

236.

Yuan, H., Xu, G., Yao, Z., Jia, J., and Zhang, Y. (2018). “Imputation of missing data intime series for air pollutants using long short-term memory recurrent neural networks.” *Proceedings 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 1293–1300.

Zhang, J., Li, D., and Wang, Y. (2020). “Predicting uniaxial compressive strength of oil palm shell concrete using a hybrid artificial intelligence model.” *Journal of Building Engineering*, 30, 101282.

Zheng, Z., Saxena, N., Mishra, K. K., and Sangaiah, A. K. (2018). “Guided dynamic particle swarm optimization for optimizing digital image watermarking in industry applications.” *Future Generation Computer Systems*, 88, 92–106.

Appendix A. CLP Factors of Dataset

Table A1. CLP factors

		Factors	Linguistic Descriptors	Scale of Measure
Activity-Level	1	Crew size	small, average, large	Integer (Total number of crew members)
	2	Craftsperson education	Elementary (1), High School (2), Technical (3), College (4), University (5)	Categorical (Most frequent category)
	3	Craftsperson on job training	poor, fair, good	Real number (No. trainings attended x Duration of Training, hrs)
	4	Craftsperson technical training	poor, fair, good	Real number (No. trainings attended x Duration of Training, hrs)
	5	Crew composition	poor, fair, good	Proportion (Ratio journeyman to apprentice to helper) (1 JR / 2 AP)
	6	Crew experience (seniority)	poor, fair, good	Real number (Crew average years of experience)
	7	Number of languages spoken	low, medium, high	Integer (Number of languages spoken, total for a crew)
	8	Co-operation among craftsperson	poor, fair, good	1–5 Predetermined rating
	9	Treatment of craftsperson by foreman	poor, fair, good	1–5 Predetermined rating
	10	Craftsperson motivation	low, average, high	1–5 rating
	11	Craftsperson fatigue	low, average, high	Real number (Total worked hours per week to Regular work hour per week)
	12	Craftsperson trust in foreman	poor, fair, good	1–5 Predetermined rating
	13	Team spirit of crew	poor, fair, good	1–5 Predetermined rating
	14	Level of absenteeism	low, medium, high	Percentage (% average number of absent crew members to total crew size, daily average)
	15	Crew turnover	low, medium, high	Turnover rate (% of crew)
	16	Discontinuity in crew makeup	small, medium, large	Real number (Average occurrence of crew member change)
	17	Level of interruption and disruption	low, medium, high	Integer (Number of interruption and disruption per day)
	18	Fairness of work assignment	poor, fair, good	1–5 Predetermined rating

19	Crew participation in foreman decision-making process	Without explanation (1), Joint (2), With (3)	Categorical (Decision Type)
20	Crew flexibility	low, average, high	1–5 rating
21	Job site orientation program	No (0), Yes (1)	Categorical
22	Job security	poor, fair, good	Integer (Average length of unemployment period, months)
23	Availability of craftsperson	poor, fair, good	Integer (Average number of unmet labour demand per crew for a given task)
24	Availability of task materials	poor, fair, good	Real number (Average waiting time for getting materials, manhours)
25	Quality of task materials	poor, fair, good	1–5 Predetermined rating
26	Material unloading practices	poor, fair, good	Real Number (average unloading time, min.)
27	Material movement practices (horizontal)	poor, fair, good	Real Number (average distance, m)
28	Material movement practices (vertical)	poor, fair, good	Real Number (average distance, m)
29	Availability of work equipment (crane, forklift)	poor, fair, good	1–5 rating
30	Availability of transport equipment (man lift)	poor, fair, good	1–5 rating
31	Equipment breakdown	infrequent, frequent, very frequent	Integer (Equipment Type and Average number of breakdown occurrence per week)
32	Availability of tools	poor, fair, good	Real number (Average waiting time, min.)
33	Sharing of tools	low, average, high	Real number (Number of crews sharing a tool)
34	Quality of tools	poor, fair, good	Real Number (Average no. of tool breakdown per week)
35	Misplacement of tools	infrequent, frequent, very frequent	Real Number (Average no. of misplacement per day)
36	Availability of electric power	poor, fair, good	Real number (Average waiting time, min)
37	Availability of extension cords	poor, fair, good	Real number (Average waiting time, min)
38	Complexity of task	low, average, high	1–5 Predetermined rating
39	Repetitiveness of task	low, medium, high	Real number (ratio of identical work tasks qty to the total work task qty)

	40	Total work volume	small, medium, large	Real number (Approved quantity for construction)
	41	Level of rework	low, average, high	Real number (Construction Filed Rework Index)
	42	Frequency of rework	infrequent, frequent, very frequent	Real number (No. of rework occurrence per scope of work)
	43	Task change orders – Extent	low, average, high	Real number (Ratio of approved total volume of change order to total work volume)
	44	Task change orders – Frequency	few, some, many	Real number (No. of occurrence per scope of work)
	45	Working condition (noise)	low, average, high	1–5 Predetermined rating
	46	Working condition (dust and fumes)	low, average, high	1–5 Predetermined rating
	47	Location of work scope (distance)	very close, close, far	Real number (distance, m)
	48	Location of work scope (elevation)	very close, close, far	Real number (distance, m)
	49	Congestion of work area	low, average, high	Real number (ratio of actual peak manpower to actual average manpower)
	50	Cleanliness of work area	poor, fair, good	Integer (Number of cleaning operations per day)
	51	Foreman skill and responsibility	poor, fair, good	1–5 rating
	52	Fairness in performance review of crew by foreman	poor, fair, good	1–5 Predetermined rating
	53	Change of foremen	infrequent, frequent, very frequent	Turnover rate (No. of turnovers per month)
	54	Span of control	low, medium, high	Integer (Average total number of crews per foreman)
	55	Response rate with RFI's	poor, fair, good	Real number (Response time, hrs)
	56	Concrete placement technique	Pump (1), Crane and Bucket (2), Direct chute (3)	Categorical
	57	Structural element	Columns (1), Footings (2), Grade Beams (3), Pile Caps (4), Slabs (5), Walls (6)	Categorical
Project-Level	58	Change in design drawings	infrequent, frequent, very frequent	Real number (Ratio of number of changed drawings to total number of drawings per discipline)

	59	Change in specifications	infrequent, frequent, very frequent	Real number (Ratio of number of changed specifications to total number of specification clauses on specific scope)
	60	Changes in contract conditions	infrequent, frequent, very frequent	Real number (Ratio of number of contract conditions changes to total number of contract clauses on specific scope)
	61	Lack of information	infrequent, frequent, very frequent	Real number (Number of RFI's per month per discipline)
	62	Approval for building permit	poor, fair, good	Real number (average process time for work or permit approval, months)
	63	Year of construction (to identify relation)	Year	Integer (Year of Construction)
	64	Project level rework	infrequent, frequent, very frequent	Real number (Project Overall CFRI)
	65	Project level change order	low, average, high	Real number (Ratio approved total cost of change order overall project to original approved project cost)
	66	Weather (temperature)	low, medium, high	Real number (˚C)
	67	Weather (precipitation)	low, medium, high	Real number (mm)
	68	Weather (humidity)	low, medium, high	Real number (%)
	69	Weather (wind speed)	low, medium, high	Real number (km/hr)
	70	Variability of weather	low, medium, high	1–5 rating
	71	Ground conditions	poor, fair, good	1–5 Predetermined rating
	72	Site congestion	low, medium, high	Real number (Ratio free site space to total site area)
	73	Width of site access	low, medium, high	Real number (Width of access, m)
	74	Queue time to access site	low, medium, high	Real number (Average queue time to access time, minutes)
	75	Project work times	poor, fair, good	1–5 rating
	76	Owner staff on site	low, average, high	Integer (Total number of owner staff on site)
	77	Approval of shop drawings and sample materials	poor, fair, good	Real number (Average time taken to approve, days)
	78	Support and administrative staff	poor, fair, good	Real number (Ratio of support to technical staff)
	79	Level of paper work for work approval	low, medium, high	1–5 rating

	80	Treatment of foremen by superintendent and project manager	poor, fair, good	1–5 Predetermined rating
	81	Uniformity of work rules by superintendent	poor, fair, good	1–5 Predetermined rating
	82	Availability of labour	low, medium, high	Real number (Unmet labour requirement, for the given trade)
	83	Labour disputes (legal cases between a worker on a project)	low, medium, high	Real number (Average number of cases per project)
	84	Project cost control	poor, fair, good	1–5 rating
	85	Labour productivity measurement practice	poor, fair, good	1–5 Predetermined rating
	86	Quality audits	low, average, high	Real number (Number of inspections per month)
	87	Inspection delay	poor, fair, good	Real number (Average delay for inspection, min)
	88	Interference	poor, fair, good	Real number (Average number of interruption due to interference)
	89	Out of sequence inspection or survey work	poor, fair, good	Real number (Number of occurrence per week)
	90	Project safety plan execution	poor, fair, good	1–5 rating
	91	Safety training	poor, fair, good	Real number (No. trainings attended x Duration of Training, hrs)
	92	Safety inspections	low, average, high	Real number (Number of inspections per month)
	93	Safety audits	low, average, high	Real number (Number of audits per month)
	94	Safety incidents	low, average, high	1–5 Predetermined rating
	95	Equipment/property damage	infrequent, frequent, very frequent	Integer (Number of reported equipment/property damage incident per month)
	96	Safety incident investigation	poor, fair, good	1 - 5 rating
	97	Project safety administration and reporting	poor, fair, good	1–5 Predetermined rating
	98	Risk monitoring and control	poor, fair, good	1–5 Predetermined rating
	99	Crisis management	poor, fair, good	1–5 Predetermined rating
	100	Communication between different trades	poor, fair, good	1–5 Predetermined rating

	101	Availability of communication devices	poor, fair, good	Real number (ratio of communication radio to number of crews, %)
	102	Hiring practices (open shop)	poor, fair, good	1–5 Predetermined rating
	103	Project team development	poor, fair, good	1–5 rating
	104	Project team closeout	poor, fair, good	1–5 rating
	105	Project environmental assurance	poor, fair, good	1–5 Predetermined rating
	106	Environmental audits	low, average, high	Real number (Number of inspections per month)
	107	Sorting of waste materials	poor, fair, good	1–3 Predetermined rating
	108	Project environmental control	poor, fair, good	1–5 Predetermined rating
Global	109	Oil price	low, average, high	Real number (Dollar/barrel)
	110	Oil price fluctuation	low, average, high	Real number (Weekly price change, %)
	111	Natural gas price	low, average, high	Real number (Dollar/GJ)
	112	Natural gas fluctuation	low, average, high	Real number (Weekly price change, %)