www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

CONCEPTUAL ESTIMATION OF CONSTRUCTION DURATION AND COST OF PUBLIC HIGHWAY PROJECTS

SUBMITTED: July 2021
REVISED: May 2022
PUBLISHED: July 2022
EDITOR: Esther Obonyo
DOI: 10.36680/j.itcon.2022.029

Basma Mohamed, PhD Student,
Department of Building, Civil, and Environmental Engineering, Concordia University, Montreal, QC, Canada
Structural Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt
basma.sayed@cu.edu.eg

Osama Moselhi, Professor,
Director of the Centre for Innovation in Construction and Infrastructure Engineering and Management (CICIEM),
Department of Building, Civil, and Environmental Engineering, Concordia University, Montreal, QC, Canada
osama.moselhi@concordia.ca

SUMMARY: *State Highway Agencies (SHAs) and Departments of Transportation (DOTs) allocate their limited resources to thousands of competing projects in multi-year transportation programs using expert judgement for the expected construction costs and durations. Such estimates overlook influencing parameters known in the planning phase and the importance of building reliable databases to support decision making. Meanwhile, it is possible to generate meaningful predictions in early stages of project development based on historical data gathering and analysis. The present research introduces a newly developed method for conceptual cost and duration estimation for public highway projects utilizing an ensemble of machine learning (ML) models and data collected for projects completed between 2004 and 2015 (roads, bridges, and drainage projects). Unlike previous studies, the proposed method includes project parameters that affect construction durations and costs and were not studied simultaneously before. The parameters considered are facility type, project scope, highway type, length, width, location, level of technical complexity, and new parameters pertinent to payment and procurement methods. The developed method was tested using 29 and 56 randomly selected projects, and the results yielded a Mean Absolute Percentage Error (MAPE) of 7.4% and 4.5% for the duration and cost, respectively, which are lower than the estimation errors of methods reported in recent literature. Additionally, the generalization abilities were assessed by the Mann-Whitney test, and the developed method is found to successfully handle diverse projects. Thus, machine learning models can assist agencies in the review process of competing projects from a high-level management perspective to ultimately develop better management execution programs.*

KEYWORDS: *Cost Estimation, Database, Duration Estimation, Highway projects, Machine Learning, Planning Phase, Transportation Programs.*

REFERENCE: *Basma Mohamed, Osama Moselhi (2022). Conceptual estimation of construction duration and cost of public highway projects. Journal of Information Technology in Construction (ITcon), Vol. 27, pg. 595-618, DOI: 10.36680/j.itcon.2022.029*

1. INTRODUCTION

The construction industry provides jobs for almost 7% of the world's working population (Pan and Zhang, 2021). Despite this economic importance, on average, only 2.5% of construction firms complete all their projects successfully without delays and/or cost overruns (Betz, 2018). This can be traced back to the uncertain, complex, and dynamic nature of construction activities as well as the interdependency between various risk factors (Gondia et al, 2020). Studies also indicate that the reason for such failure is the unrealistic and inconsistent estimation of duration and/or cost in the pre-design phase (Halpin et al, 2017). However, it is illogical to spend manhours and resources to conduct detailed estimates for all proposed projects at that early stage.

State highway agencies (SHAs) are constrained by the limited available resources and the deteriorating state of existing infrastructure (Antoine et al, 2019) and should therefore avoid the repercussions of budget and schedule problems. To start with, both cost overrun and cost underrun are problematic (Zhang et al, 2017). Cost underruns may cause a suboptimum number of projects to be executed wasting the agency resources, while overruns may lead the agency to overspend and borrow from the next year's budget to compensate the shortage, thereby impairing the agency's credibility (Zhang et al, 2017). Additionally, schedule delays can potentially lead to disrupting the economic growth when road users must wait for the projects' completion longer than necessary (El-Maaty et al, 2017). However challenging, having reliable duration and cost estimates in early project stages can mitigate these community problems and their consequences for the involved stakeholders (Le et L, 2021). For instance, the awarding agency can use these reliable estimates to rationalise the process of project prioritization, making financial decisions, assessing work zone and traffic impacts, better coordinating with third-party stakeholders such as neighbouring businesses, and planning public safety (Qiao et al, 2019, Le et al, 2021). Hence, the present study aims to address the question of how we can improve the estimation of construction duration and cost in the planning stage of highway projects?

2. LITERATURE REVIEW

2.1 Cost and Duration Estimation

Cost and duration estimation has been the focus of many studies in both the academia and the industry. State Highway Agencies develop project duration estimates only after the planning phase is completed using bar charts, critical path method, and the estimated cost method (Nevett et al, 2021). In addition, there were several attempts by agencies and departments to develop customized duration estimation tools. For instance, Louisiana Department of Transportation adopts production rate-based duration estimation based on the quantities of each activity (Nevett et al, 2021). These tools, however, were reported to have poor accuracy (Zhai et al, 2016). For instance, one study estimated the average prediction error of two such tools as greater than 200% across various projects (Taylor et al, 2012, Abdel-Raheem et al, 2018). One reason for this could be overlooking the influential factors impacting highways durations and costs.

Several studies have attempted to model the relationship between highway projects features and their durations in early stages (Nevett et al, 2021, Titirla and Aretoulis 2019, Son et al, 2019, Okere, 2019, Nani et al, 2017, Pesko et al, 2017 and others). Pesko et al, (2017) concluded that integrating the estimation of duration and cost can compromise the significance of input data, leading to reduction of estimating accuracy compared to that based on separate estimation. Nevett et al, (2021) used multiple linear regression to estimate the duration of 1,500 highway projects in the early planning stage based on cost data, traffic volume, terrain type, project condition, project size, and construction quantities. Their model was found to have a Mean Absolute Percentage Error (MAPE) of 44%. This level of accuracy has not been achieved by other researchers for a model that includes various project sizes and types. Pesko et al, (2017) conceptually estimated the duration of 166 urban roads in the tender offer by using construction material quantities, cost category, and work distribution across the different categories. Their Artificial Neural Network obtained a MAPE of 26.26%, and the Support Vector Machine Model (SVM) achieved a slight improvement with 22.77% MAPE.

On the other hand, the American Association of State Highway and Transportation Officials (AASHTO, 2013) explains two procedures to early cost estimation in its practical guide: conceptual estimating and bid-based estimating. Conceptual estimates support the early planning of projects especially with limited available information. An example is the unit cost (center lane mile or square foot of area), which is calculated using past projects data and other project related factors (AASHTO, 2013). Bid-based estimates rely on bid items, project quantities, and historical data after its adjustment for the current project conditions (AASHTO, 2013). The present research is concerned with the first type, which is the conceptual estimation of highway construction cost using known data at this stage. It is also known as the Top-Down Estimation method (Karaca et al, 2020), which only relies on the high-level project features and does not require detailed information about the project, such as the items quantities.

Compared to duration estimation studies, more research was done for cost estimation in the pre-construction phase of projects (Elmouslami, 2020, Dehghan et al, 2020, Meharie and Shaik, 2020, Juszczyk, 2020, Tijanic et al, 2020, Karaca et al, 2020, Aretoulis, 2019, Meharie et al, 2019, Cao et al, 2018, Zhang et al, 2017, Gardner et al, 2017, Chou et al, 2015). Juszczk (2020) applied SVM on bridge projects completed between 2005 and 2018 to estimate their cost in the pre-construction phase by delivery method, length, width, and structural properties, and the optimum model had a MAPE of 10.94%. Further, Tijanic et al, (2020) employed ANN to estimate the cost of 57 road sections completed in the last 20 years in Croatia in the initial design phase based on their length, width, and contracted duration with a 13% MAPE. Cao et al, (2018) utilized an ensemble of Machine Learning models to estimate the cost of 1,400 unit-price resurfacing highway projects, and their MAPE was found to be 7.56%.

Tables 1 and 2 highlight knowledge gaps in recent literature of duration and cost estimation for highway projects. Generally, these highway-related studies are far less than those conducted for building construction projects and contain several limitations. For example, highway studies that achieve high accuracies are either applicable to certain project types or develop separate models for different project types. This can hinder their suitability to meet highway agencies' need for including different project types in their transportation programs.

TABLE 1: Conceptual Duration Estimation Studies

Study	**Techniques**	**Parameters used**	**Findings**
Nevett et al, 2021	Multiple Linear Regression on 1,500 completed projects	Cost Project Characteristics AADT: Average Annual Daily Traffic Terrain Type Project Size Construction Quantities	Estimation Error (44%). No sensitivity analysis to determine the most influential variables on the duration. Planning quantities may not necessarily be precisely determined during the conceptual phase.
Titirla and Aretoulis, 2019	Neural Networks and Correlation Analysis on 37 highway projects in Greece	Initial cost Project Scope: Length No. of lanes No. of technical projects Embankment Existence of bridges, geotechnical projects, and landfills.	Mean Squared error with a value of 6.96E-06 No sensitivity analysis to determine the most influential variables on the duration.
Son et al, 2019	Multiple Linear Regression on 623 completed highway projects in Dallas, USA	Cost Type of Work County Size (Population)	Estimation error (35.89%). No feature selection method or sensitivity analysis to determine the most influential variables on the duration. Results applicable only to low-bid design-bid-build projects, in one geographic region
Nani et al, 2017	Relative Importance Index and Stepwise Regression Artificial Neural Networks using 30 completed bridge projects in Ghana	Bridge Span Bridge Weight Bridge Lane Excavation In situ Concrete Reinforcement Formwork Site Clearance	Estimation Error (25%-26%) No sensitivity analysis to determine the most influential variables on the duration. Material quantities may not necessarily be precisely determined during the conceptual phase. Very Small Sample Size Results applicable only to bridge projects. Some of the parameters of the model are only applicable to developing countries
Pesko et al, 2017	Artificial Neural Networks and Support Vector Machine on 166 projects of basic construction works and/or reconstruction of urban roads	Quantity of construction materials Distribution of the work across the different work categories Project type	Estimation Error (22.77%-26.26%) No feature selection method to determine the most influential variables on the duration. Planning quantities may not necessarily be precisely determined during the conceptual phase.

TABLE 2: Conceptual Cost Estimation Studies

Study	Techniques	Parameters used	Findings
Meharie and Shaik, 2020	Random Forest, Artificial Neural Network, and Support Vector Machine using 74 completed highway projects	Project length Number of bridges Inflation rate Project scope Terrain type Project location	The best model was found to be the Random Forest model with RMSE 0.9579, followed by the ANN model with RMSE 1.1802, and finally the SVM model with RMSE 1.2569. No sensitivity analysis to determine the most influential variables on the cost.
Juszczyk, 2020	Support Vector Machine using road bridges, rail bridges, and animal bridges completed in Poland between 2005 and 2018.	Type of Bridge Method of Delivery Length, Width Number of Spans Structural Parameters of bridge	Estimation Error (10.94%). No feature selection method to determine the most influential variables on the cost. Results applicable only to bridge projects.
Tijanic et al, 2020	Artificial Neural Network on 57 road sections completed in the last 20 years in The Republic of Croatia	Project Scope Length and Width of Road Construction Duration	Estimation Error (13%) No feature selection method or sensitivity analysis to determine the most influential variables on the cost. Results applicable only to road projects.
Aretoulis, 2019	Artificial Neural Network on 20 highway projects.	Culvert Construction Project Type Number of Lanes Tunnel Construction Initial Duration District Relocating Utilities Bridge Construction Project Length Earthmoving Cost Paving Cost Surfacing Cost Signs Cost Electromechanical Cost Technical Works Cost Geomorphology	The optimum neural network model produced a mean squared error with a value of 7.68544E-05 No sensitivity analysis to determine the most influential variables on the cost.
Meharie et al, 2019	Multiple Linear Regression with no dataset	Project Type Project Complexity Project Location Project Size Site Topography Bridge Type Number of Bridges Existence of groundwater Soil Type Inflation Rate Project Duration	The developed method was neither applied nor validated. No sensitivity analysis to determine the most influential variables on the cost.
Pesko et al, 2017	Artificial Neural Networks and Support Vector Machine on 166 projects of basic construction works and/or reconstruction of urban roads	Quantity of construction materials Distribution of the work across the different work categories Project type	Estimation Error (7.06%-25.38%) No feature selection method to determine the most influential variables on the cost. Planning quantities may not necessarily be precisely determined during the conceptual phase.
Zhang et al, 2017	Least Absolute Shrinkage and Selection Operator (LASSO) Regularized Regression on 741 completed highway projects	CPPRs (contractors' past performance ratings) Contract days Weather days Length of the project Number of lanes Consumer Price Index (CPI) Prime Lending Rate (PLR) Producer Price Index (PPI)	Estimation Error (7.1%) No feature selection method or sensitivity analysis to determine the most influential variables on the cost. Limited to resurfacing projects in Florida. More focused on the financial aspects than the current project characteristics.

Tables 1 and 2 also confirm that cost estimation studies outnumber the duration studies. Further, highway agencies develop duration estimates only after the planning phase is completed, unlike cost estimates, which are prepared in early phases and revised later on. Additionally, having more data is expected to lead to better accuracy.

However, no recent study is comprehensive and combines project characteristics (Sharma et al, 2021, Meharie and Shaik, 2020, Son et al, 2019), project scope (Qiao et al, 2019, Okere, 2019, Juszczyk, 2020), cost and schedule information (Sharma et al, 2021, Titirla and Aretoulis, 2019), level of technical complexity (Karaca et al, 2020, Gardner et al, 2017) and expected risk profile for different project types in one robust general model with acceptable accuracy. Some studies even use item quantities for duration and/or cost prediction in the conceptual phase when they may not be precisely determined (Nevett et al. 2021; Pesko et al. 2017; Nani et al. 2017). Besides, some studies rely solely on datasets that comprise factors whose probability and/or impact are determined subjectively by experts in questionnaires (Le et al, 2021). This could be attributed to the shortage of historical project data impeding quantifiable data-driven studies (Le et al, 2021). Many studies also lack sensitivity analysis and/or feature selection to identify the most influential project variables on duration and/or cost of highway projects.

Therefore, highway agencies need a comprehensive way to successfully manage their long-term strategic programs with respect to both schedule and budget. This warrants the need for an estimation method that does not require activity durations or work sequences, since they are not known at this early stage. Such method also ought to achieve high prediction accuracy and at the same time, require small effort in development and update. Consequently, the estimators working hours can be freed to only work on the detailed estimation of projects that were initially pre-approved by the highway agency, thereby capitalizing on more potentially successful business opportunities.

Unlike methods reported in the literature, the present study aims to provide a simple objective estimation method that models the relationship between a comprehensive set of high-level project parameters that are known in the planning and development phase, and the project duration and cost. This method can assist highway agencies in development of their own estimates for various project types such as: roads, bridges, and drainage projects. The developed method expands on previous studies by including project parameters that were not studied simultaneously before, adding new parameters, combining a range of project sizes and types in one model, and determining parameters that best predict construction duration and cost in the planning phase. However, any time or cost components corresponding to project phases before construction, such as the design preparation, are not included in the scope of this study.

2.2 Modelling Approaches

Estimation of construction duration and cost before the design completion can be accomplished either by analogous/comparative/conceptual methods or statistical/predictive methods (Son et al, 2019). Comparative estimation methods rely on analogy with similar past projects to estimate the duration and/or cost of the current project during its early stages (Son et al, 2019). This is done via combining historical information and expert judgement by the project manager (Hashemi et al, 2020). However, this method is questionable, as it assumes a linear relationship between the project cost and its basic design variables such as the location, size, type, and capacity of the structural components (Chakraborty et al, 2020) and doesn't take into account data known at this stage such as payment and procurement methods. Predictive or Statistical estimation methods are also used before design completion relying on the project parameters known at this stage rather than activities relationships and quantities (Son et al, 2019). Hence, they provide higher reliability and accuracy than conceptual estimation (Son et al, 2019). Statistical methods mainly seek to model the relationship between the input variables (project features from a dataset of past projects) and the output (duration and/or cost) (Hashemi et al, 2020). Statistical methods include parametric and non-parametric modelling. The main difference is that parametric modelling such as regression analysis requires the definition of the mathematical form first before conducting any analysis, where both the dependent and independent factors must follow the parameters established for the modelling relationship: linear or non-linear (El-Sawah and Moselh,i 2014). This is contrary to non-parametric modelling which relies on fitting an unknown equation to the dependent variable and its predictors, and the number and nature of the parameters are determined from the data and not fixed in advance (Juszczyk, 2017).

Various studies have employed regression analysis to estimate the cost and/or duration of highway projects in the conceptual phase (Nevett et al, 2021, Son et al, 2019, Zhang et al, 2017, Bayram, 2017, Zhai et al, 2016, Czarnigowska and Sobotka, 2014). Even though regression analysis is easier to interpret, is faster when learning from the data, and requires less training data than non-parametric techniques (Kotu and Deshpande,, 2014), it was illustrated that it behaves poorly in the estimation of durations and cost in the conceptual phase of highway projects (low accuracy). Furthermore, the few studies which successfully obtain high accuracy such as those conducted by Zhang et al, (2017) and Czarnigowska and Sobotka (2014) were applicable to certain project types: resurfacing projects and public roads, respectively. This can be attributed to the fact that it is a highly constrained process which requires the priori assumption of the type of relationship between the predictors and the dependent variable beforehand (Juszczyk, 2017). Such assumption hinders proper modelling of the interaction between the different project variables, and this does not suit the dynamic nature of construction, especially during the conceptual phase that contains many uncertainties (El-Sawah and Moselhi, 2014). Further, the relationships between the input parameters and the output may exhibit a mix of linear and non-linear relationships, and thus, they may not be properly captured by regression analysis (Matel et al, 2019). Additionally, the residuals of linear regression models must meet certain assumptions. They should be independent and normally or approximately normally distributed. Their mean should be zero, and their variance must be a constant (Zhang et al, 2017). Usually, the data gathered from the field does not meet these assumptions. Consequently, the regression models become susceptible to multicollinearity, which means the presence of strong correlations among the independent variables leading to unstable model predictions (Elmousalami, 2020).

On the other hand, non-parametric estimators are capable of learning by detecting the implicit relationships between the input parameters and the output without having initial assumptions. This is particularly useful for the conceptual phase of construction projects when there is limited prior knowledge (Juszczyk, 2017). As a result, they are slower to train than parametric models, as they require more training data to estimate the mapping function. They also have the risk of overfitting, so they require scrutiny in their development and interpretation. However, these limitations are offset by their flexibility and being powerful modelers in the sense of making no prior assumptions. They can also approximate non-linear relationships with high accuracy and analyze raw data independent of their distribution or even type (Elmousalami, 2020). Hence, they have gained popularity in the construction engineering and management field owing to their ability to solve complex problems successfully (Mensah et al, 2016). In fact, non-parametric modelling has outperformed regression analysis in several studies (Singh and Chauhan, 2009, Petruseva et al, 2013, Chou et al, 2015). For instance, Chou et al, (2015) used regression analysis and ANN for estimating the cost of bridge projects, and the ANN model outperformed the regression model with a MAPE of 7.56%. Consequently, non-parametric modeling techniques are used in the present research since high accuracy and robustness are cornerstones in prediction applications that involve resource planning and significant financial decisions (Tijanic et al, 2019).

2.3 Machine Learning Algorithms

Artificial intelligence is a fundamental cornerstone in improving the way a construction project is planned and performed (Pan and Zhang, 2021). Machine learning is a branch of artificial intelligence that allows computers to learn and derive meaning from their prior experiences. i.e., raw input data, without having to be directly programmed (Sharma et al, 2021). This is corresponding to humans' perception abilities that can handle complex problems intentionally and adaptively (Pan and Zhang, 2021). Currently, the construction industry is rapidly increasing AI investments, most of which is directed towards machine learning related applications that can learn huge and diverse amounts of data and make smart decisions, accordingly. Such innovation is considered reasonable given the known fact that the construction field is very rich and data intensive. Consequently, machine learning is well-suited for the goal of this study. It can be categorized into supervised and unsupervised learning based on the type of data presented to the model (Sharma et al, 2021). Supervised learning/Guided learning/ Learning with a teacher provides the desired output as an input to the machine, so the model learns from labelled data unlike unsupervised learning (Sharma et al, 2021). Since the available data consists of project parameters rather than activities relationships and quantities, the studied system is too complex to have known or traceable patterns. Therefore, supervised ML is best suited to handle it since it focuses on learning from implicit data patterns (Kotu and Deshpande, 2014).

ML models are often prone to overfitting and hyper-parameter selection issues (Elmousalami, 2021). Overfitting occurs when the training algorithm corresponds too closely with the training data and fails to predict unseen future records reliably. Hyper-parameter selection is crucial for optimizing the learning process of a ML algorithm. As a result, it is generally advised to create more than one predictive model (Elmousalami, 2021). In the current study, Artificial Neural Networks (ANN), Support vector Machines (SVM), and Random Forest (RF) algorithms were selected, as they are commonly used in construction research domain (Elmousalami, 2021, Meharie and Shaik, 2020, Juszczyk, 2020, Titirla and Aretoulis, 2019, Pesko et al, 2017). They were subsequently combined in an ensemble to benefit from their collective strengths and balance their weaknesses.

2.3.1 Artificial Neural Networks (ANN)

A neural network contains multiple layers: input, output, and several hidden layers, as seen in Fig. 1. The input layer obtains information from the outside world, just like the human brain, and passes it onto the hidden layers. ANNs have been repeatedly cited in the literature. For example, they have been used for the conceptual estimation of highway projects cost especially when limited information is available (Meharie and Shaik, 2020, Tijanic et al, 2019, Karaca et al, 2020, Aretoulis, 2019, Cao et al, 2018, Gardner et al, 2017, Chou et al, 2015). Additionally, they have been applied to estimate the duration of highway projects in their early stages, though not as widely as cost (Titirla and Aretoulis, 2019, Nani et al, 2017, Pesko et al, 2017, Mensah et al, 2016). ANN has gained popularity in several disciplines owing to its superiority in comparison with other AI techniques for several reasons (Sharma et al, 2021). It can handle large amounts of complex non-linear data with fast real time performance without having to conform to any assumptions such as linearity or normality (Kotu and Deshpande, 2014, Mensah et al, 2016), which is highly suited for construction data gathered from the field. It can also achieve high predictive accuracy by capturing complex associations and patterns even if the causal relationships are ambiguous (Karaca et al, 2020). Finally, it can effectively handle small sized datasets with independent attributes, like the dataset included herein (Kotu and Deshpande, 2014).

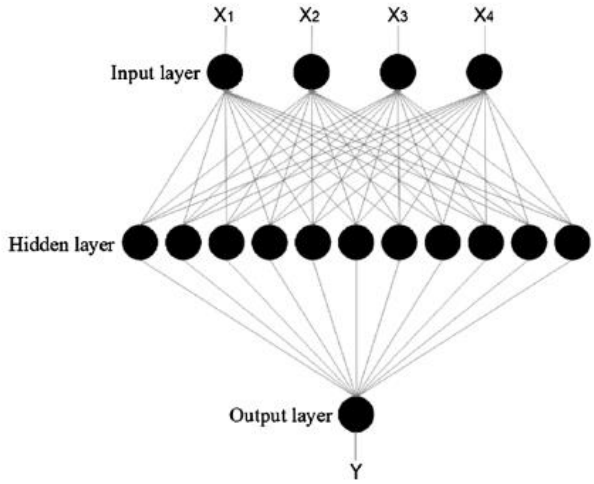

FIG. 1: General Structure of ANN (Tijanic et al, 2019)

2.3.2 Support Vector Machine (SVM):

Support Vector Regression (SVR) creates a decision surface by mapping input vectors into a high-dimensional (or infinite-dimensional) feature space, where regression is performed (Wauters and Vanhoucke, 2014). Since the relationship between the multidimensional input vector x and the output y is mostly unknown and possibly non-linear, this mapping operation is required (Kotu and Deshpande, 2014). There have been multiple successful applications of SVM. Pesko et al, (2017) reported that SVMs surpassed ANNs in the conceptual estimation of both durations and costs in the tender stage of urban roads. Juszczyk (2020) obtained a MAPE of 10.94% in the prediction of construction costs of bridge projects in their early stages using SVM. On the other hand, SVM can often have slow computation times because of the higher dimensional mapping, which can become challenging in machine learning problems with many attributes. However, they are widely applied as they seek to reduce the generalization error considering both the training error and a confidence level, unlike other techniques that only minimize the training error (Zhang al, 2019), and they also perform exceptionally well with both linear and non-linear data (Wauters and Vanhoucke, 2014). Usually, construction data can include both linear and non-linear relationships, requiring algorithms that can handle such diversity easily. Additionally, their robustness and overfitting resistance against changes in the training data make SVMs one of the most versatile machine learning algorithms (Kotu and Deshpande, 2014). Finally, most machine learning models focus their learning process on the most common cases (far from the boundary), SVM prioritizes the extreme rare cases strengthening its predictive ability (Kotu and Deshpande, 2014).

2.3.3 Random Forest (RF)

Random Forest was proposed by Breiman (2001). It is a version of ensemble learning, where it consists of a set of decorrelated decision trees to avoid their frequent overfitting issues (Kotu and Deshpande, 2014). Each tree is trained using a bootstrapped sample of the training data (drawn with replacement) on a subset of the parameters that are randomly selected for each tree (Chakraborty, 2020). The final result is predicted for every new record using a majority vote in case of classification or by averaging the predictions for regression tasks, thereby improving the predictive accuracy (Chakraborty, 2020). Many studies employed Random Forest alongside other ML techniques for the estimation of construction projects cost (Cao et al, 2018, Chakraborty et al, 2020, Meharie and Shaik, 2020). Understandably, Random Forest requires extensive computation resources and time to train multiple trees and combine their outputs. It can be complex and require comprehensive training data, as it does not predict beyond the range of its previously seen data (Chakraborty, 2020). On the other hand, it offers the following advantages for data-based prediction problems: it is more stable and less susceptible to performance changes due to changes in the dataset than individual decision trees (Kotu and Deshpande, 2014), it can handle large numbers of input parameters (Cao et al, 2018), and it can effectively handle small sized datasets with independent attributes (Kotu and Deshpande, 2014). Therefore, it offers an opportunity for flexible modelling, which is convenient for construction related applications.

2.3.4 Ensemble Learning

ML algorithms suffer from generalization errors that are composed of two parts: an irreducible error (cannot be eliminated) and a reducible error (bias and variance) (Ahiaga-Dagbui and Smith 2014). The irreducible error is the noise which is by definition a random intrinsic aspect of data, and little can be done about it (Aggarwal, 2015). Bias is the average error between the actual and predicted values of the target output, while variance refers to the sensitivity of the model to different training sets (Ahiaga-Dagbui and Smith, 2014). The modelling effort should be directed towards the reducible error (Aggarwal, 2015). However, if we attempt to reduce the bias, the variance will increase and vice versa.

A bias-variance trade-off can be achieved using ensemble methods, where several models are combined to improve their overall predictive performance (Moon et al, 2020). In other words, a strong learner is built using several weak learners. It is sometimes compared to consulting a "committee of several experts", the individual models in this case, before making a final decision (Ahiaga-Dagbui and Smith, 2014). This enables the end user to collectively benefit from their strengths and balance their weaknesses as well as reduce the risk of overfitting (Aggarwal, 2015). An ensemble model effectively deals with high-dimensionality issues, complex data structures, small sample size, and missing data (Elmousalami, 2020). On the downside, the presence of multiple models may increase the complexity, thereby reducing the interpretability (Elmousalami, 2020). Nonetheless, the numerous benefits including the improved predictive accuracy and reduced risk for overfitting certainly outweigh the interpretability issues, especially in applications that are more concerned with making accurate financial decisions.

There are many studies in various fields in which the ensemble learners surpassed the individual base models. Cao et al. (2018) developed an ensemble of Extreme Gradient Boosting, ANN, and RF for the prediction of unit-price bids of resurfacing projects. The proposed ensemble learning model performed much better than the baseline models with MAPE of 7.56%. Ahiaga-Dagbui and Smith (2014) developed a pre-contract cost estimation model using non-parametric bootstrapping and ensemble learning of ANNs for water infrastructure projects. The ensemble models were superior to the bootstrapping model, where 92% of the validation predictions were found within +/- 10% of the actual cost and 77% within +/- 5% of the actual cost. Zou et al. (2020) applied an ensemble model comprising a gradient boosting tree and a deep neural network to predict the travel time of highways. These studies indicate better performance of ensemble models than individual base models.

Ensemble models can be developed either by (Moon et al, 2020):

1. Voting which is mainly averaging of various models' predictions on one level or
2. Stacking where the predictions of multiple models on the same dataset are combined in two levels. The predictions produced from the Level-0 Base Models are used to build a new Level-1 Meta Model, which makes the final predictions on the testing dataset.

Fig. 2 explains the difference between voting and stacking ensemble models.

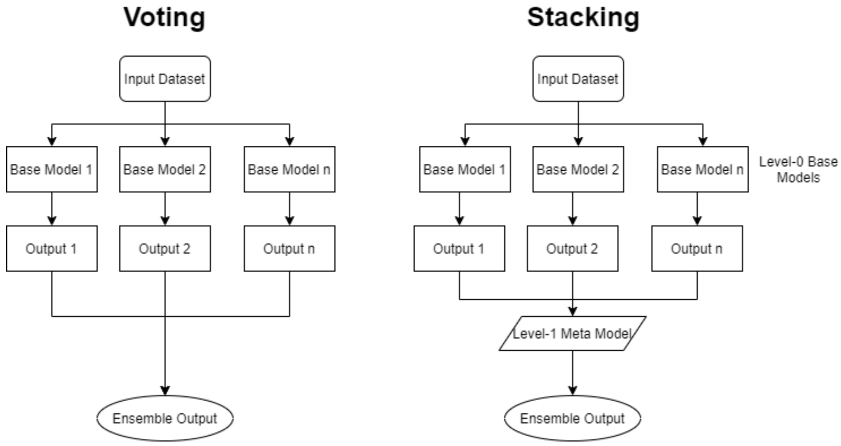

FIG 2: Voting vs Stacking Ensemble Models (adapted from Moon et al, 2020)

3. DEVELOPED METHOD

A newly developed method is presented for conducting efficient conceptual estimates of cost and duration for highway projects. The method encompasses four main steps/phases (Fig. 3):

1. Identifying the factors impacting the costs and durations of this class of projects
2. Collecting and organizing data
3. Training the Machine Learning Models
4. Testing and Verifying the Machine Learning Models

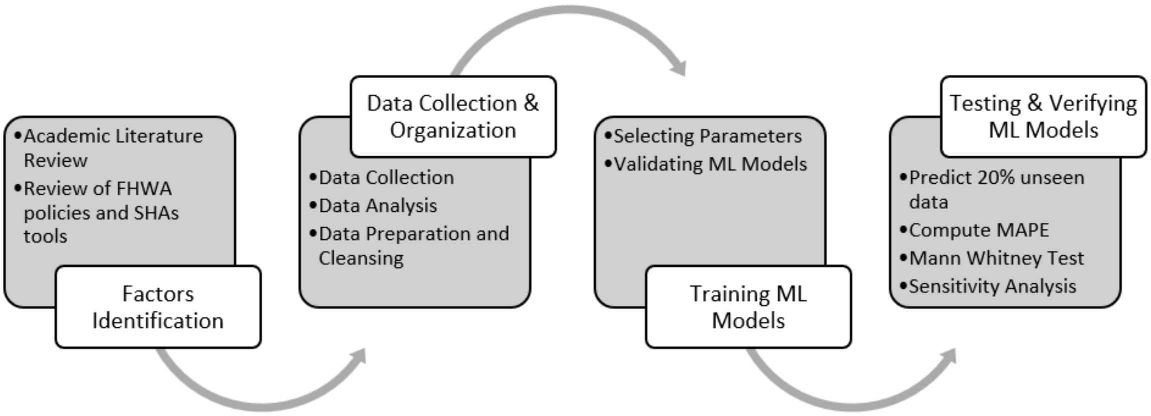

FIG. 3: Developed Method

3.1 Factors Identification

The most recent and relevant research studies were examined to extract the factors influencing the conceptual estimation of construction duration and cost of highway projects. More specifically, the studies selected were those conducted between 2017 and 2022 and concerned with the estimation of construction duration and cost in the planning phase of highway projects when not much detailed information is available. This selection process was done by the help of Google Scholar search engine to ensure a comprehensive database search. For instance, the extensive review study conducted by Sharma et al, (2021) identified various necessary input parameters impacting highway projects preliminary durations and costs including location, procurement, payment, facility type, project scope, highway type, and project technical complexity. Le et al, (2021) also presented an extensive literature review to highlight the most influencing factors on a public roadway project duration: project technical complexity and project risk profile.

The current study aims to expand upon the study conducted by Nevett et al, (2021) for estimating highway projects construction duration using multiple linear regression, which obtained an error of 44%. This level of accuracy has not been achieved by other researchers for a model that includes various project sizes and types. Selecting the parameters to be included in the conceptual estimation models was based on the recent literature. The project location was one of the most cited factors for duration estimation by Sharma et al, (2021), Le et al, (2021), and Son et al, (2019). The facility type also impacts the schedule of a highway project (Sharma et al, 2021, Le et al, 2021, Qiao et al, 2019). The reviewed cost estimation literature included facility type (Sharma et al, 2021, Tijanic et al, 2019, Aretoulis, 2019), highway type (Sharma et al, 2021, Juszcczk, 2020, and Karaca et al, 2020), and expected technical complexity level (Sharma et al, 2021, Karaca et al, 2020, Meharie et al, 2019, Gardner et al, 2017). Furthermore, the project location was employed in various cost estimation studies (Sharma et al, 2021, Meharie and Shaik, 2020, Aretoulis, 2019).

Tables 3 and 4 include the parameters used in the developed method, their types, and source. The target output variables in the collected dataset refer to the total project duration in working days including design and construction and the construction cost in USD including change orders.

TABLE 3: Estimation Parameters for Construction Duration

Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Parameter Type**	**Possible Values**
1. Project Agency	●	●	●					Categoric	US State (Location)
2. Facility Type	●	●		●				Numeric	Roads, Bridges, Drainage, and Intelligent Transportation System (ITS) as % of Total Cost
3. Project Scope	●				●			Numeric	New Construction/Expansion, Rehabilitation/Reconstruction, and Resurfacing/Renewal
4. Highway Type	●		●					Numeric	Rural Interstate, Urban Interstate, Rural Primary, Urban Primary, and Rural Secondary.
5. Length in Kilometers						●	●	Numeric	Kilometers
6. Number of Lanes						●	●	Numeric	Number
7. Technical Complexity	●	●						Numeric	Non-complex, Moderately Complex, and Most Complex

TABLE 4: Estimation Parameters for Construction Cost

Factor	**1**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Parameter Type**	**Possible Values**
1. Project Agency (Location)	●	●	●				●			Categoric	US State (Location)
2. Facility Type	●						●	●		Numeric	Roads, Bridges, Drainage, and Intelligent Transportation System (ITS) as % of Total Cost
3. Project Scope	●		●							Numeric	New Construction/Expansion, Rehabilitation/Reconstruction, and Resurfacing/Renewal
4. Highway Type	●			●	●					Numeric	Rural Interstate, Urban Interstate, Rural Primary, Urban Primary, and Rural Secondary.
5. Length in Kilometers				●						Numeric	Kilometers
6. Number of Lanes				●						Numeric	Number
7. Technical Complexity	●				●	●			●	Numeric	Non-complex, Moderately Complex, and Most Complex

Sources for Construction Duration & Cost:

1. Sharma et al. 2021
2. Le et al. 2021
3. Son et al. 2019
4. Qiao et al. 2019
5. Okere 2019
6. Titirla and Aretoulis 2019
7. Nani et al. 2017
8. Liang et al. 2021
9. Meharie and Shaik 2020
10. Juszczyk 2020
11. Karaca et al. 2020
12. Meharie et al. 2019
13. Aretoulis 2019
14. Tijanic et al. 2019
15. Gardner et al. 2017

Moreover, the present study includes other project attributes which have not been included in duration and/or cost estimation studies to date:

1. Payment Method: Unit price, Lump Sum, Guaranteed Maximum Price, and Cost Reimbursable (Categoric Parameter)
2. Procurement Method: Low Bid, Best Value, Qualifications-based, and (A+B) Cost + Time (Categoric Parameter)

3.2 Data Collection and Organization

3.2.1 Data Collection

Usually, data is hard to obtain, as the construction industry is still largely deficient in recording and publishing projects data. The present study was developed using data from the literature, as a part of the research study "Project Delivery Methods' Change-Order Types and Magnitudes Experienced in Highway Construction" (Alleman et al, 2020) and other studies conducted by the same authors such as (Antoine et al, 2019). Fortunately, the authors were generous enough to provide us with their original dataset. The collected data includes but is not limited to the project name, project ID, project agency, project location, facility type, project scope, highway type, construction cost, engineering services cost, construction duration, change orders values, encountered delays, lessons learned, project risk profile, project delivery method, payment method, procurement method, and the project technical complexity. A sample project information is found in Appendix A. The projects sizes range between two million and fifty million USD. The collected data comprised 291 US highway construction projects completed between 2004 and 2015 from state and federal highway transportation agencies: duration data was available for 146 projects, and cost data was available for 284 projects. Projects were completed in several states including Florida, Utah, Arizona, Colorado, Oregon, and Maine among others (FHWA, 2018). An adjustment for inflation was done by the FHWA National Highway Construction Cost Index to ensure a fair comparison of the projects' costs at the same time by converting them to equivalent values in June 2015.

The purpose of the original study was to examine the relationships between highway projects delivery methods and the types and magnitudes of encountered change orders (Alleman et al, 2020). Initially, this data was collected by the authors from state and federal highway agencies through mining cost and schedule performance datasets and distributing a questionnaire survey to projects' representatives to obtain additional project performance data. This was part of a big first-of-a-kind FHWA national study in 2018, and at the time, it was considered one of the biggest empirical highway construction datasets (FHWA, 2018). The data collection process was completed under rigorous quality control measures for 18 months, and there was no need to merge data from various sources for any one project (Antoine et al, 2019). All the collected data can be found in the final FHWA project report (FHWA, 2018). The current research utilizes the variables relevant to its scope of work, and these variables are sometimes referred to as: parameters, factors, and attributes.

3.2.2 Preliminary Data Analysis and Preparation

In the present study, data analysis is done to ensure that the dataset is correctly representative of US public highway projects and hence corroborate the applicability of the concluded insights in the study. For example, the percentages of Design-Bid-Build (DBB), Design-Build (DB), and Construction Management/General Contractor (CM/GC) are 40%, 48%, 12%, respectively, of the collected projects sample, which is representative of US highway projects according to FHWA (2016). Contrary to popular belief, agencies use alternative contracting methods (DB and CM/GC) on all project sizes not only larger projects (FHWA, 2018). It was also noted that projects twice as large were completed in half the time by DB and CM/GC in comparison with DBB (FHWA, 2018). Regarding the procurement method, almost 60% of the dataset were based on the lowest bid, which is particularly common with government-owned projects.

The data preparation process involves the following steps:

1. Handling Missing Data and Outliers: The collected dataset had no missing values within the variables selected for the scope of the current research. On the other hand, the modified z-score, which is suitable for smaller datasets than z-score, was applied to eliminate the outliers using Equation 1 (Abdelaty et al, 2020). The data points whose modified z-score is more than 3.5 are eliminated from further analysis. Thus, five projects were excluded from the dataset based on this value.

   $M_i = (0.6745\ (X_i\text{-}X'))/MAD$ …. Equation 1

   where M_i: Modified z-score, Xi: Observed Value, X': Median, and MAD: Absolute Difference between median and observed value.

2. Encoding categorical data: Most machine learning models can only handle numeric attributes and not nominal data. Hence, categorical attributes such as the location are transformed into several columns of 0 and 1 (dummy variables)

3. Splitting into training and testing datasets: In this research, the dataset was divided into 80% for training the model and 20% for testing it, using a stratified sampling procedure to ensure the same distribution of the output variables in the two sub-samples (Kotu and Deshpande, 2014).

4. Feature Scaling: Scaling means to reduce all data values (both input and output) to the same order of magnitude (Pesko et al, 2017). Hence, all data will have equal significance in the modelling process in contrast to having a smaller attribute be overlooked by a larger attribute like the cost (some values reach millions). The method applied here is the Zero-Mean normalization, which entails shifting the entire dataset vertically so that their average is equal to zero. It was selected because of its simplicity and suitability for all kinds of distributions in the dataset (Pesko et al, 2017).

3.2.3 Size of Dataset

An important aspect in developing machine learning models is the size of the data set used for the training and testing processes. Small sized datasets can adversely affect the reliability of the developed models through overfitting, where the machine learning model memorizes the specific patterns of the training records instead of establishing a generalized structure from the records and/or generating highly varied performance indices among the test folds (Kim et al, 2008). On the other hand, large datasets can often contain noise, and this necessitates the use of noise modeling and outlier analysis techniques (Gondia et al, 2020).

It should be noted that the present model satisfies the minimum sample size reported by Elmousalami et al, (2020) and Kass and Tinsley (1979) for construction estimation modelling which is between 5 and 10 times the number of all variables. Hence, for the construction cost and duration models, the minimum sample size would be between 5x9= 45 and 10x9 = 90. The gathered data included 146 and 284 data points for the duration and cost models, respectively satisfying the minimum data requirements. Nevertheless, it can still be considered a relatively small dataset in the realm of machine learning based predictive models. This small size of datasets in most studies is attributed to the challenges of obtaining confidential information owned by governmental agencies about highway projects (Karaca et al, 2020). Hence, the selection of the adopted ML models was limited to those techniques that are well-suited to handle such consideration with a demonstrated history of satisfactory performance. As a final note, even though the limited number of project records can somehow impede the model generalizability, this is highly dependent upon having access to more data. The developed method can be applied to larger datasets where the required data is available to produce customized models for highway projects cost and duration prediction.

3.2.4 Analysis of Lessons Learned Reports Text Data

Even though the information included in the lessons learned and bidding documents is classified as unstructured data in the world of machine learning, it was found quite useful when converted into structured data in the format used by ML algorithms. Further, the inferences made from this data can be quite beneficial in predictive applications, like the study at hand. The dataset had more complete information about the encountered delays than cost overruns. Table 5 shows the words and word pairs that were found to be associated with the three delay levels:

1. Less than 30% (Minor)
2. Between 30% and 60% (Moderate)
3. More than 60% (High)

For instance, for projects with high delay, the associated words reveal some encounters and activities that can be challenging to estimate or perform such as: "government shutdown", "utility relocation", "disputes", "cranes productivity", "labor productivity", and "traffic control". Hence, having established systems in place to deal proactively with possible claims, disputes, and productivity issues can help avoid high levels of schedule delays in future projects. It is also shown that such projects often experienced uncertainties such as "unknown quantity" or "unpredicted condition". Thus, encountering such words in future projects will prompt the planners to request clarification in advance during the bidding stage. Projects with moderate delays are shown to relate to other word pairings that can explain why they had a lower level of delay such as: "aggressive schedule" and "reduced impact". Another interesting inference is that moderate delay projects included "value engineering" and "completion incentive". The idea behind value engineering is to encourage the construction contractor, with a monetary incentive, to make suggestions to improve the design constructability, cost effectiveness, and smoothness of installation, since contractors are typically more experienced in this regard (Halpin and Senior, 2017). There are fewer word pairs for minor delay projects than the previous two categories, and the most frequent words are found in Table 5. Hence, the high and moderate delay groups can be differentiated from each other and from the other group, as their risk factors were more predictive and exclusive. This unstructured text data can provide the user with invaluable insights, that interpret the past projects' experiences and can be helpful to plan for future projects as well to avoid unnecessary delays.

TABLE 5: Text Analysis Results

	High	**Moderate**	**Minor**
Word/Word Pair	government_shutdown	damage	expedite_project
	utility_relocation	aggressive_schedule	traditional_design
	traffic_control	value_engineering	allowed_overlap
	Replace	commit	ahead_schedule
	crane_productivity	completion_incentive	
	Assess	negotiate	
	unknown_quantity	reduced_impact	
	unpredicted_condition	additional	
	Disputes	minimize_impact	
	labor_productivity	accelerate	

3.3 Training Machine Learning Models

Training ML models comprises selecting optimum parameters and validation to ensure a satisfactory and reliable performance for the developed models.

3.3.1 Selecting the parameters of the ML Models

The success of any ML model highly depends on the tuning of its hyperparameters (Wauters and Vanhoucke, 2014). Selecting values for these hyperparameters was based on a) following the best practices from previous studies to ensure practicality, simplicity, and low computation requirements, b) suitability for the task at hand, and c) Grid Search. Grid Search Optimization is a wrapper method which uses a feedback loop, namely the cross-validation technique, to iteratively select values for the parameters of a ML model to customize its performance to the dataset (Kotu and Deshpande, 2014). Grid search optimization was selected because of its low requirements of computation time and resources, as it is well-suited for conducting small quick searches among a group of hyperparameter values, that are generally known to provide a satisfactory performance (Wauters and Vanhoucke, 2014). 80% of the gathered data was randomly selected for the optimization process, which corresponds to 117 and 228 projects for the estimation of duration and cost, respectively.

3.3.2 Validating the ML Models

The k-fold cross validation method is known for its reliability in the assessment of both the training error and the generalization error and the optimization of the bias-variance trade-off (Gondia et al, 2020). In this method, the entire data set is randomly split into k distinct and almost equal number of folds, where k is a positive integer. One fold is set aside for testing, and the remaining k-1 folds are used for training the model. The testing fold is known as the validation sub-set, and it is mainly used to fine tune the model parameters. This process is repeated k times. The predictive performance of the model is evaluated by averaging the selected prediction metrics across the different folds (Wauters and Vanhoucke, 2014). According to the literature, the best value for k is 10 to ensure optimal computation time and estimation of error (Kohavi, 1995). The grid search is conducted by repeating this process for the different combinations of the model hyperparameters, and the combination that achieves the best score across the k folds is selected (Wauters and Vanhoucke, 2014). The main advantage in this technique is that the entire data set is used for training and testing the model, and each example is used once for testing. Further, this eliminates the probability of overfitting and strengthens the generalization ability of the developed model (Wauters and Vanhoucke, 2014).

The next sections explain the parameter selection and validation processes of the individual ML models and the formulation of the ensemble learning model which will be subsequently tested.

3.3.3 Artificial Neural Networks (ANN):

ANNs have a few parameters to be selected. The selected network architecture is the Multi-Layer Perceptron (MLP), as it is commonly used for regression applications (Pesko et al, 2017). It consists of an input layer which receives the task to be processed, an arbitrary number of hidden layers which process the input data, and an output layer which performs the required regression task (Abirami and Chitra, 2020). Other architectures exist for ANN such as the convolutional neural network which has more sophisticated uses like image processing, object detection, optical character recognition, natural language processing, and sound (Abirami and Chitra, 2020). Another type of ANN is the recurrent neural network which is more suited for modeling sequence data like sound or time series sensor data than MLP (Abirami and Chitra, 2020). Moreover, the back propagation algorithm is selected for practicality owing to its simplicity, excellent generalization capability, and lower computation requirements in comparison to General Regression and Radial Basis Function Neural Networks (Tijanic et al, 2019). The optimum number of hidden layers has been confirmed by many results in different engineering fields to be not more than two (Huang and Lipmann, 1988, Pesko et al, 2017). In fact, it is generally recommended to begin with only one hidden layer in the grid search (Hyari et al, 2016). In the present study, the grid search was done with 1 and 2 hidden layers to minimize the computation speed and requirements. Selecting the optimum number of neurons in the hidden layers is necessary to avoid the two extreme cases of: too many hidden neurons leading to overfitting, or insufficient neurons causing the omission of basic functions (Pesko et al, 2017). In this study, the number of neurons in the hidden layers was set to 1,2,3,5,7,9 in the grid search. Fig. 4 shows the variation of prediction error with the number of neurons in the hidden layers for the duration ensemble model. The activation function translates the given input or set of inputs into an output. The options for regression are either the rectifier function or hyperbolic tangent (Pesko et al, 2017). The learning rate is responsible for the amount of updating the model according to the prediction error, thereby controlling the training speed. The grid search is done between 0.01 and 1 (Kotu and Deshpande, 2014). 200 epochs were repeated for every combination in the grid search to eventually obtain the least prediction error. Table 6 shows the optimum values for the duration and cost ANN models.

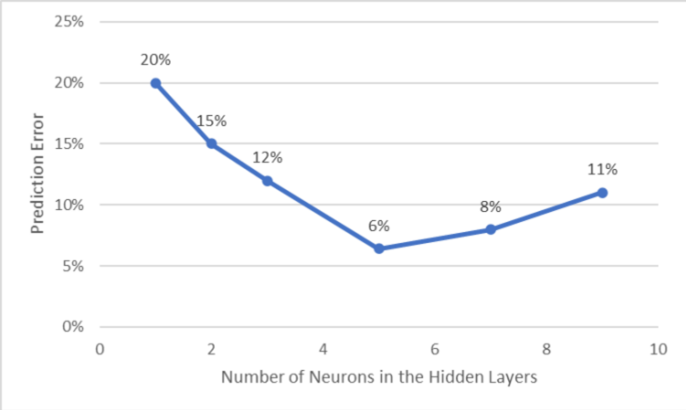

FIG. 4: Number of Neurons in the Hidden Layers vs Prediction Error

TABLE 6: Optimum Values for the Ensemble Learning Models

	No. of Hidden Layers	No. of Neurons in Hidden Layers	Activation Function	Learning rate
Construction Duration	1	5	Hyperbolic Tangent	0.01
Construction Cost	1	3	Hyperbolic Tangent	0.035

3.3.4 Support Vector Machine (SVM):

SVMs also have several parameters to be selected. Kernel functions transfer the nonlinear spaces into linear ones (Kotu and Deshpande, 2014). The possible types include Gaussian Radial Basis Kernel Function (RBF), Sigmoid Kernel, and Polynomial Kernel. In this study, the RBF Kernel was selected due to its demonstrated ability to map non-linear relationships and having little numerical difficulties (Wauters and Vanhoucke, 2014). Fig. 5 shows an example of a developed Support Vector Regression Model showing the remaining hyperparameters. In order to obtain an optimum performance, we need to minimize Equation 2:

$$0.5 \mid w \mid\mid^2 + C\sum(\varepsilon - \varepsilon^*) \quad \text{.... Equation 2}$$

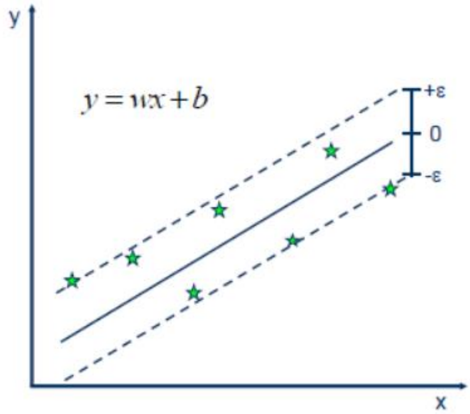

FIG. 5: Support Vector Regression

We must also select the SVM complexity constant "C" that controls the optimum trade-off between good training and generalization behavior (Wauters and Vanhoucke, 2014). Values that are too high create softer boundaries, while values that are too low create more rigid boundaries (Kotu and Deshpande, 2014). Sequences of C = (2^{-5}, 2^{-3},…, 2^{13}, 2^{15}) are tried. The Epsilon parameter defines the width of the insensitive tube around the true values of the target variable. The prediction error is discarded if the difference between the actual and predicted values is less than ε (Juszczyk, 2020). Its range in this study ranged between 0.05 and 0.2 with step of 0.05. Last but not least, the kernel gamma for the radial basis function is assumed as the inverse of the number of inputs (1/9=0.111).

Table 7 shows the optimum combinations that resulted in the lowest error, which will be used for the validation sub-set.

TABLE 7: Optimum Values for the Ensemble Learning Models

	C	**Epsilon**
Construction Duration	2^7	0.1
Construction Cost	2^5	0.15

3.3.5 Random Forest (RF):

Even though Random Forests do not have many parameters that require optimization in a grid search, they are generally computationally challenging and time-consuming. The main reason for this would be the number of trees. As rule of thumb, the greatest performance boost occurs in the first 250 trees or so (Kotu and Deshpande, 2014). The selected splitting criterion is the least square for numeric attributes (Kotu and Deshpande, 2014). The sub-set ratio of randomly chosen attributes to test must be selected carefully to maintain the balance between bias and variance. A small value will likely reduce the variance at the cost of increasing the bias. However, this can be avoided if the data has high quality and few noisy features. On the other hand, for more noisy data, a larger value is encouraged to include more worthy variables (Kotu and Deshpande, 2014). Its range in this study ranged between 0.125 and 0.5 with step of 0.05. The optimum ratio was found to be 0.2 and 0.3 for the duration and cost models, respectively. Finally, the size of the Bootstrapped Dataset parameter is often neglected because the bootstrapped dataset of the individual trees, which is sampled with replacement, will be completely different from the complete training set, even if they have the same size. As a result, each tree can be trained with a bootstrapped sample of the same size as the training data (Kotu and Deshpande, 2014).

3.3.6 Ensemble Machine Learning Model

Generally, different ML models have different sources of bias and variance. Simple models such as SVM and shallow decision trees have high bias, since they make a lot of assumptions about the prediction boundaries (Aggarwal, 2015). More complex models such as ANN and RF have high variance due to their tendency to overfit the data (Aggarwal, 2015). As a result, it is crucial to select the component predictors to optimize the bias-variance trade-off. In the present study, two ensemble learning methods are adopted to combine the individual trained ML models:

1. Voting: Averaging of predictions on one level from ANN, SVM, and RF.
2. Stacking: Combining the predictions from multiple models using two levels:
    a. Level-0 Base Models: ANN & RF
    b. Level-1 Model (Meta Model): SVM

3.4 Testing and Verifying Machine Learning Models

Models of complex systems are usually simplifications of reality because it is rather difficult to fully represent the characteristics of the actual systems virtually. Additionally, the construction industry usually considers ML models to be black boxes and practitioners are reluctant to adopt them until they assess their predictive performance, their ability to extrapolate on unseen data, and the degree of importance of the features (Tijanic et al, 2019). Thus, performance evaluation of such models is a fundamental step to reduce the uncertainty and ensure the reliability of the output results (Balci, 1994). The evaluation process includes two steps: Testing and Verification.

1. Testing: Using empirical data, that wasn't exposed to the model during its training, to test whether it accurately represents the system being modelled in a realistic and acceptable way (Balci, 1994), thereby assessing the model's predictive performance through the hold-out method and the ability to generalize its predictions via the Mann Whitney U test.
2. Verification: Making sure that inputs are correctly replicated as outputs by testing the functionality of the model using some of its parameters (Balci, 1994) and measuring the degree of importance of the model's variables through sensitivity analysis.

3.4.1 Testing

The training and testing of the predictive models were done on different subsets of the data to ensure an unbiased evaluation using the hold out method. The hold-out method helps the modeler assess the generalization capabilities of a ML model by testing its ability to provide satisfactory results through applying the prediction to a sub-set of the data that was not known to the model during its development (training) (Pesko et al, 2017). The hold out method is very common in developing machine learning models, where the entire data set is randomly split into two portions: 60-80% and 40-20% for training and testing the models, respectively (Gondia et al, 2020). In this study, the dataset was divided into 80% and 20%. 80% of the gathered data was used in the training of the ML models, and the remaining 20% are used for the testing. The developed method was tested using 29 and 56 randomly selected projects, for the duration and cost, respectively. The evaluation technique used in this study is the Mean Absolute Percentage Error (MAPE), which compares the actual and predicted values of the target output. It is calculated as per Equation 3, and its classification ranges are found in Table 8.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{y\ actual\ i - y\ pred\ i}{y\ pred\ i}\ x\ 100\ \ldots.\ Equation\ 3$$

TABLE 8: Classification Ranges (Elmousalami 2020)

MAPE Value Range	Classification of ML model
<10%	Excellent
10% - 20%	Good
20%-50%	Acceptable
>50%	Inaccurate

Tables 9 and 10 show the MAPE values for the ML individual models. It is evident that the individual models had MAPE values ranging between 10% and 20% indicating that they are ‘good’ predictors. The voting ensemble models obtained MAPE values lower than 10% (6.4% for duration and 4.3% for cost), and the stacking ensemble models obtained MAPE values lower than 10% (7.4% for duration and 4.5% for cost) improving the predictive ability to ‘excellent’. Hence, combining a group of ‘good learners’ led to the formation of an ‘excellent learner’. Such results outperform those of similar recent studies conducted to estimate highway construction durations and achieved a MAPE of 44% using parametric estimation (Nevett et al, 2021). One reason could be the necessity of having normally distributed data, a condition which can not always be satisfied by the data collected from the construction field. Moreover, the cost estimation results of the developed method are superior to those of Tijanic et al, (2019) which obtained a MAPE of 13% for road projects and Juszczyk (2020) whose study had a MAPE of 11% for bridge projects using non-parametric estimation. In addition, the proposed method estimation error constitutes an improvement above the customized agencies tools whose error was found to be 200% in some studies (Abdel-Raheem et al, 2018). Further, the obtained results are in conformity with the literature finding that prediction of construction duration is more challenging than cost (Pesko et al, 2017, Yeom et al, 2018). This is demonstrated by the results where the error percentages of ensemble duration models are higher than those of ensemble cost models. The model was further tested with 40% and 50% testing data, and it was found that the MAPE fluctuated around +/- 0.05% only.

TABLE 9: Duration Models

	MAPE
SVM	15.4%
ANN	16%
RF	15.2%
Voting	6.4%
Stacking	7.4%

TABLE 10: Cost Models

	MAPE
SVM	16.7%
ANN	14.2%
RF	14.8%
Voting	4.3%
Stacking	4.5%

3.4.2 Verification (Sensitivity Analysis)

Changing the input data can reveal the sensitivity of the model to changes in such input parameters (Balci, 1994). This is specifically important for complex models (Kermanshachi and Rouhanizadeh, 2019). For verification purposes, the input variables will be divided as follows:

1. Group 0: Project Scope, Highway Type, Length, Number of Lanes, and Location
2. Group 1: Facility Type
3. Group 2: Procurement and Payment Methods
4. Group 3: Technical Complexity

Changing one element at a time (OAT) to see if it affects the performance is one of the most realistic ways of conducting sensitivity analysis and has been commonly favored by modelers in a number of fields (Kermanshachi and Rouhanizadeh, 2019). This method entails modifying one of the input variables while leaving the others unchanged, and then repeating the procedure with all other inputs in the same way. The sensitivity is measured by measuring changes in the system's output, which is a rational method since any observed shift in the output is attributable to the moving variable's changes. The same folds of projects for cross validation were used in these new runs to obtain better insight about the various model parameters. This process was repeated for each of the four variable groups, and the models were run 20 times for each parameter.

Table 11 shows the conducted model runs, the resulting Mean Absolute Percentage Error, and the included model parameters/features for the construction duration model.

Table 12 shows the conducted model runs, the resulting Mean Absolute Percentage Error, and the included model parameters/features for the construction cost model.

TABLE 11: Sensitivity Analysis Results for Duration Models

Attribute	**MAPE (%)**	**% Improvement**	**Rank**
Group 0	18.5%		
Group 0 + Group 1	15.3%	3.2%	2
Group 0 + Group 1 + Group 2	9.7%	5.6%	1
Group 0 + Group 1 + Group 2 + Group 3	7.4%	2.3%	3

TABLE 12: Sensitivity Analysis Results for Cost Models

Attribute	**MAPE (%)**	**% Improvement**	**Rank**
Group 0	16%		
Group 0 + Group 1	9.85%	6.15%	1
Group 0 + Group 1 + Group 2	6.5%	3.35%	2
Group 0 + Group 1 + Group 2 + Group 3	4.5%	2%	3

The base model was conducted using the variables from Group 0 to be easily comparable to the studies conducted in the field (Nevett et al, 2021, Meharie and Shaik, 2020, Juszczyk, 2020). These variables reflect the project characteristics. The additional runs were conducted by adding 1 group at a time (Facility Type, Payment and Procurement Methods, and Level of Technical complexity). From the results, it is evident that the duration is more sensitive to the choice of payment and procurement methods, followed by facility type and level of technical complexity. On the other hand, the impact of input parameters on the output value is not the same for cost. Thus, it was found that cost is impacted by facility type, payment and procurement methods, and finally level of technical complexity. This analysis addresses gaps in the literature of related studies, as most of the studies in the literature lack a sensitivity analysis (Nevett et al, 2021, Meharie and Shaik, 2020, Juszczyk, 2020, Tijanic et al, 2020, Aretoulis, 2019, Meharie et al, 2019, Son et al, 2019, Pesko et al, 2017). Consequently, the output of this research can guide researchers and industry practitioners acquire knowledge about the key project variables that were established to impact highway projects durations and costs. These variables were not included in similar studies before (payment and procurement methods), included in only few studies (level of technical complexity), or overlooked in many studies (facility type).

3.5 Assessing Applicability to a Variety of Highway Projects

The Mann Whitney U Test is conducted to evaluate the generalization abilities of the developed models: whether the distributions of the actual and estimated values for the dataset were the same. The Mann Whitney U test, also known as the Wilcoxon Rank Sum Test or the Mann Whitney Wilcoxon Test, is used to determine if two samples are likely to come from the same population (i.e., that the two populations have the same shape) (Son et al, 2019). The test combines two groups, ranks their values, and compares the rankings of each group's values separately to see if one is smaller than the other (Son et al, 2019).

The following are the null (H_0) and alternative hypotheses (H_a) for the two-tailed Mann-Whitney test.

1. H_0: Both the actual and estimated values have the same distributions (i.e., neither distribution is smaller).
2. H_a: The actual and estimated values do not have the same distributions (i.e., either distribution is smaller).

The Mann Whitney U test was conducted on 25 random sub-samples of 15 projects from the validation subset of the dataset. The test was repeated to ensure that various types and sizes were included and to have a number of projects that is large enough to form a distribution. The null hypothesis could not be discarded at the significance level of 0.05 since the p-value was 0.2144 for construction duration and 0.5982 for construction cost. The test was also performed on the entire validation subset (58 projects), yielding the same results. This confirms the proposed framework's generalizability to the various types of highway projects included in the dataset (roads, bridges, and drainage projects) since the distributions of the actual and estimated values were the same. This is contrary to previous studies that obtained comparable accuracies in duration and/or cost estimation for highway projects in the pre-design phase and were limited to certain types of highway projects. Examples include Juszczyk (2020) who estimated the costs of bridge projects with an error of 10.94% and Tijanic et al (2020) who employed ANN to assess the costs of road sections and obtained an error of 13%. Finally, the present study outperforms the study conducted by Nevett et al (2021) which included a variety of highway project types and sizes in one model but was only limited to duration estimation and an error of 44%.

4. CONCLUSION

Proper gathering and mining of construction projects' data can create useful corporate value and facilitate evidence-based decisions. Therefore, this study aims to provide highway governmental agencies with a fast and reliable method to estimate the construction durations and costs of different highway projects included in their multi-year transportation programs. This is achieved using non-parametric statistical estimation utilizing various types of project-related data documented for past projects, which were identified from recent research studies and industrial reports. It enables estimation of highway projects duration prior to design completion, which has received little attention from researchers compared to cost estimating. Additionally, the construction industry has not fully exploited the capabilities of machine learning in making use of the abundance of historical projects' data for estimation purposes, and the present research aims to bridge this gap. Such data is far from being fully utilized and is either lost or seen in limited perspective. Ultimately, this can prevent knowledge loss and lead to re-using it to sustainably build capacity and maintain efficient resource allocation among new construction and rehabilitation projects, while addressing the drawbacks of the current literature in this respect. Testing efforts revealed that the proposed method has excellent predictive accuracy over previously developed methods in the literature: 7.4% for duration and 4.5% for cost. Additionally, the performed sensitivity analysis reveals that the estimation of highway projects construction durations and costs is most sensitive to the constructed facility type as well as the choice of payment and procurement methods, which have not been included in duration and/or cost estimation studies to date. The developed method's applicability to a wide range of highway projects was also proved by the Mann-Whitney U Test. Furthermore, interesting insights were concluded from analyzing the projects' Lessons Learned Reports. For example, highway projects that experienced more than 60% delays were found to be associated with events pertaining to "government shutdown, utility relocation, disputes, crane productivity, and unpredicted condition". What's more, the reports of projects with lower level of delays (30%-60%) had other occurrences such as "aggressive schedule, value engineering, and completion incentive", and reports of projects with less than 30% delays had word associations such as "allowed overlap". These occurrences clarify their improved performance. Such insights can interpret the past projects' experiences and improve future projects planning to avoid preventable setbacks in highway projects execution.

Notwithstanding, the presented method holds certain constraints and opportunities for future research. For instance, as with any data-driven methodology, the specific properties and conditions associated with the analyzed dataset means that the resulting numeric values are not necessarily transferable to other cases. However, the collected data contains enough information to create meaningful and reliable estimations and be applicable to a variety of highway projects. Further, the results herein are determined by the average performance of many projects. Any one project can outperform or underperform the average. Additionally, the proposed prediction method does not model the uncertainty involved in multiple projects selection in a SHA's long-term program. For instance, similar projects could cause managerial problems in case of competition for resources. Concerning the covered projects, the facility types are limited to roads, bridges and drainage projects, and the project delivery methods are limited to DB, DBB and CM/GC delivery methods. Other delivery methods such as Public Private Partnership (PPP) and Integrated Project Delivery (IPD) are not included. Future research studies can also address conducting a sensitivity analysis to assess the duration estimation risk associated with the accuracy of the cost estimate, since the analysis is conducted in early project stages, where the cost estimate is subject to change.

5. ACKNOWLEDGEMENTS

The authors are immensely grateful for Mr. Douglas Alleman (email: doal7544exc@colorado.edu), PhD Candidate at the University of Colorado, Boulder for his invaluable support and providing the dataset which facilitated the implementation of this research study.

REFERENCES

AASHTO. (2013). Practical guide to cost estimating. The American Association of State Highway and Transportation Officials, Washington, DC, USA.

Abdelaty, A., Shrestha, K. J., & Jeong, H. D. (2020). Estimating preconstruction services for bridge design projects. *Journal of Management in Engineering*, Vol. 36 No. 4, 04020034.

Abdel-Raheem, M., A. Neira, C. Torres-Cantu, and X. Wang. 2018a. Initial assessment of the current contract time determination systems used by TxDOT. In Proc., *International Conference on Transportation and Development 2018*, 373–380. Reston, VA: ASCE.

Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. *Advances in Computers*, Vol. 117, No. 1, 339-368.

Aggarwal, C. C. (2015). Data mining: the textbook. Springer, New York, USA.

Alleman, D., Antoine, A. L., Stanford, M. S., & Molenaar, K. R. (2020). Project Delivery Methods' Change-Order Types and Magnitudes Experienced in Highway Construction. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, Vol. 12, No.2, 04520006.

Antoine, A. L., Alleman, D., & Molenaar, K. R. (2019). Examination of project duration, project intensity, and timing of cost certainty in highway project delivery methods. *Journal of Management in Engineering*, Vol. 35 No. 1, 04018049.

Aretoulis, G. N. (2019). Neural network models for actual cost prediction in Greek public highway projects. *International Journal of Project Organisation and Management*, Vol. 11, No. 1, 41-64.

Balci, O., 1994, Validation, verification, and testing techniques throughout the life cycle of a simulation study, *Annals of operations research*, Vol. 53, No. 1, 121-173.

Bayram, S. (2017). Duration prediction models for construction projects: In terms of cost or physical characteristics?. *KSCE Journal of Civil Engineering*, Vol. 21, No. 6, 2049-2060.

Betz, J. 2018. 27+ impressive project management statistics in 2019. Accessed June 24, 2021. https://learn.g2.com/project-management-statistics.

Bini, S.A. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *Journal of Anthroplasty*. 2018, Vol. 33, 2358–2361.

Breiman, L. (2001). Random forests. *Machine learning*, Vol. 45, No. 1, 5-32.

Cao, Y., Ashuri, B., & Baek, M. (2018). Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. *Journal of Computing in Civil Engineering*, *Vol. 32*, *No. 5*, 04018043.

Colorado Department of Transportation (CDOT). 2013. Design bulletin: Construction phasing. Colorado Department of Transportation, Denver, Colorado, USA.

Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, Vol. 46, 101201.

Chou, J. S., Lin, C. W., Pham, A. D., & Shao, J. Y. (2015). Optimized artificial intelligence models for predicting project award price. *Automation in Construction*, Vol. 54, 106-115.

Czarnigowska, A., & Sobotka, A. (2014). Estimating construction duration for public roads during the preplanning phase. *Journal of Engineering, Project, and Production Management*, Vol. 4, No. 1, 26-35.

Dehghan, R., Mortaheb, M. M., & Fathalizadeh, A. (2020). A heuristic approach to forecasting the delivery time of major project deliverables. *Practice Periodical on Structural Design and Construction*, Vol. 25 No. 2, 06020004.

Dominic D. Ahiaga-Dagbui & Simon D. Smith (2014) Dealing with construction cost overruns using data mining, *Construction Management and Economics*, Vol. 32, No. 7-8, 682-694,

El-Maaty, A. E. A., El-Kholy, A. M., & Akal, A. Y. (2017). Modeling schedule overrun and cost escalation percentages of highway projects using fuzzy approach. *Engineering, Construction and Architectural Management*, Vol. 24, No. 5, 809-827.

Elmousalami, H. H. (2020). Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis. *IEEE Transactions on Engineering Management*, Vol. 68, No. 1, 183-196.

El-Sawah, H., & Moselhi, O. (2014). Comparative study in the use of neural networks for order of magnitude cost estimating in construction. *Journal of Information Technology in Construction (ITcon)*, Vol. 19, No. 27, 462-473.

Federal Highway Administration (FHWA). 2016. Quantification of cost, benefits, and risk associated with alternative contracting methods and accelerated performance specifications. Federal Highway Administration, Washington, DC, USA.

Federal Highway Administration (FHWA). 2018. TechBriefs: Alternative contracting method performance in US highway construction. Federal Highway Administration, Washington, DC, USA.

Gardner, B. J., Gransberg, D. D., & Rueda, J. A. (2017). Stochastic conceptual cost estimating of highway projects to communicate uncertainty using bootstrap sampling. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, Vol. 3, No. 3, 05016002.

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine learning algorithms for construction projects delay risk prediction. *Journal of Construction Engineering and Management*, Vol. 146, No. 1, 04019085.

Halpin, D. W., Lucko, G., & Senior, B. A. (2017). Construction management. John Wiley & Sons, Hoboken, New Jersey, USA.

Hashemi, S. T., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences*, Vol. 2 No. 10, 1-27.

He, X., Liu, R., & Anumba, C. J. (2021). Data-Driven Insights on the Knowledge Gaps of Conceptual Cost Estimation Modeling. *Journal of Construction Engineering and Management*, Vol. 147, No. 2, 04020165.

Hyari, K. H., Al-Daraiseh, A., & El-Mashaleh, M. (2016). Conceptual cost estimation model for engineering services in public construction projects. *Journal of Management in Engineering*, Vol. 32, No. 1, 04015021.

Hyosoo Moon , Trefor P. Williams , Hyun-Soo Lee & Moonseo Park (2020) Predicting project cost overrun levels in bidding stage using ensemble learning, *Journal of Asian Architecture and Building Engineering*, Vol. 19, No. 6, 586-599,

Juszczyk, M. (2017). The challenges of nonparametric cost estimation of construction works with the use of artificial intelligence tools. *Procedia engineering*, Vol. 196, 415-422.

Juszczyk, M. (2020). On the Search of Models for Early Cost Estimates of Bridges: An SVM-Based Approach. *Buildings*, *Vol. 10, No. 1*, 2.

Karaca, I., Gransberg, D. D., & Jeong, H. D. (2020). Improving the accuracy of early cost estimates on transportation infrastructure projects. *Journal of Management in Engineering*, Vol. 36, No. 5, 04020063.

Kass, R. A., and H. E. A. Tinsley. 1979. Factor analysis. *J. Leisure Res.* Vol. 11, No. 4, 120–138.

Kermanshachi, S., & Rouhanizadeh, B. (2019, June). Sensitivity Analysis of Construction Schedule Performance Due to Increased Change Orders and Decreased Labor Productivity. *7th CSCE International Construction Specialty Conference (ICSC)* (pp. 12-15).

Kim, H., L. Soibelman, and F. Grobler. 2008. Factor selection for delay analysis using knowledge discovery in databases. *Automation in Construction*, Vol. 17, No. 5, 550–560.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence* Vol. 14, No. 2, 1137–1145.

Kotu, V., & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with RapidMiner. Morgan Kaufmann, Waltham, Massachusetts, USA.

Le, C., Wai Yaw, M., David Jeong, H., & Choi, K. (2021). Comprehensive Evaluation of Influential Factors on Public Roadway Project Contract Time. *Journal of Management in Engineering*, Vol. 37, No. 5, 04021044.

Lee, E.-B., H. Lee, and C. W. Ibbs. 2007. Productivity aspects of urban freeway rehabilitation with accelerated construction. *Journal of Construction Engineering and Management*, Vol. 133, No. 10, 798–806.

Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2019). An artificial neural network approach for cost estimation of engineering services. *International journal of construction management*, Vol. 0, No. 0, 1-14.

Meharie, M. G., & Shaik, N. (2020). Predicting Highway Construction Costs: Comparison of the Performance of Random Forest, Neural Network and Support Vector Machine Models. *Journal of Soft Computing in Civil Engineering*, Vol. 4, No. 2, 103-112.

Meharie, M. G., Gariy, Z. C. A., Ndisya Mutuku, R. N., & Mengesha, W. J. (2019). An effective approach to input variable selection for preliminary cost estimation of construction projects. *Advances in Civil Engineering*, Vol. 2019.

Mensah, I., Adjei-Kumi, T., & Nani, G. (2016). Duration determination for rural roads using the principal component analysis and artificial neural network. *Engineering, Construction and Architectural Management,* Vol. 23, No. 5, 638-656.

Nani, G., Mensah, I., & Adjei-Kumi, T. (2017). Duration estimation model for bridge construction projects in Ghana. *Journal of Engineering, Design and Technology*, Vol. 15 No. 6, 754-777.

Nevett, G., Goodrum, P. M., & Littlejohn, R. L. (2021). Understanding the Effect of Bid Quantities, Project Characteristics, and Project Locations on the Duration of Road Transportation Construction Projects during Early Stages. *Transportation Research Record*, Vol. 2675, No. 2, 121-134.

Okere, G. (2019). An Evaluation of a Predictive Conceptual Method for Contract Time Determination on Highway Projects Based on Project Types. *International Journal of Civil Engineering,* Vol. 17, No. 7, 1057-1073.

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517.

Pesko, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D., & Krklješ, M. (2017). Estimation of costs and durations of construction of urban roads using ANN and SVM. *Complexity*, Vol. 2017, 2450370.

Petruseva, S., Zujo, V. and Zileska-Pancovska, V. (2013), Neural network prediction model for construction project duration, *International Journal of Engineering Research & Technology (IJERT)*, Vol. 2 No. 11, 1646-1654.

Probst, P., & Boulesteix, A. L. (2017). To Tune or Not to Tune the Number of Trees in Random Forest. *Journal of machine Learning Research*, Vol. 18, No. 1, 6673-6690.

Qiao, Y., Labi, S., & Fricker, J. D. (2019). Hazard-based duration models for predicting actual duration of highway projects using nonparametric and parametric survival analysis. *Journal of management in engineering*, Vol. 35, No. 6, 04019024.

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2020). Machine learning model for delay risk assessment in tall building projects. *International Journal of Construction Management*, Vol. 0, No. 0, 1-10.

Sharma, S., Ahmed, S., Naseem, M., Alnumay, W. S., Singh, S., & Cho, G. H. (2021). A Survey on Applications of Artificial Intelligence for Pre-Parametric Project Cost and Soil Shear-Strength Estimation in Construction and Geotechnical Engineering. *Sensors*, Vol. *21, No.* 2, 463.

Singh, Y. and Chauhan, A.S. (2009), Neural networks in data mining, *Journal of Theoretical and Applied Information Technology*, Vol. 5 No. 6, 36-42.

Son, J., Khwaja, N., & Milligan, D. S. (2019). Planning-Phase Estimation of Construction Time for a Large Portfolio of Highway Projects. *Journal of Construction Engineering and Management*, Vol. 145, No. 4, 04019018.

Taylor, T. R., M. Brockman, D. Zhai, P. M. Goodrum, and R. Sturgill. 2012. Accuracy analysis of selected tools for estimating contract time on highway construction projects. *Proc., Construction Research Congress 2012 American Society of Civil Engineers*. Reston, VA: ASCE.

Tijanic, K., Car-Pušić, D., & Šperac, M. (2019). Cost estimation in road construction using artificial neural network. *Neural Computing and Applications*, Vol. 32, 9343-9355.

Titirla, M., & Aretoulis, G. (2019). Neural network models for actual duration of Greek highway projects. *Journal of Engineering, Design and Technology,* Vol. 17, No. 6, 1323-1339.

W. Y. Huang and R. P. Lipmann. (1988). Neural Net and Traditional Classifiers Neural Information Processing Systems, D. Z. Anderson, Ed., American Institute of Physics, New York, USA.

Wauters, M., & Vanhoucke, M. (2014). Support vector machine regression for project control forecasting. *Automation in Construction*, Vol. 47, 92-106.

Yeom, D.-J., H. M. Seo, Y.-J. Kim, C.-S. Cho, and Y. Kim. 2018. Development of an approximate construction duration prediction model during the project planning phase for general office buildings. *Journal of Civil Engineering and Management*, Vol. 24, No. 3, 238–253.

Zhai, D., Shan, Y., Sturgill, R. E., Taylor, T. R., & Goodrum, P. M. (2016). Using parametric modeling to estimate highway construction contract time. *Transportation Research Record*, Vol. 2573, No. 1, 1-9.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, Vol. 99, 238-248.

Zhang, Y., Minchin Jr, R. E., & Agdas, D. (2017). Forecasting completed cost of highway construction projects using LASSO regularized regression. *Journal of Construction Engineering and Management*, Vol. 143, No. 10, 04017071.

Zou, Z., Yang, H., & Zhu, A. X. (2020). Estimation of travel time based on ensemble method with multi-modality perspective urban big data. IEEE Access, 8, 24819-24828.

APPENDIX A

Sample Project Data used in Training:

Project ID	Hidden for privacy
Project Location	Tacoma
Facility Type	34% Roads, 25% Bridges, and 14% Drainage
Project Scope	Reconstruction
Highway Type	Urban Interstate
Construction Cost	Hidden for Privacy
Construction Duration	3,646 working days
Change Orders Value	Hidden for Privacy
Project Delivery Method	DBB
Payment Method	Unit Price
Procurement Method	Low Bid
Project Technical Complexity	1
Length in Km	100
Number of Lanes	6
Uncertainty in Geotechnical investigation (Cost)	4
Uncertainty in Geotechnical investigation (Duration)	4
Work Zone Traffic Control (Cost)	2
Work Zone Traffic Control (Duration)	2
Environmental Impacts (Cost)	2
Environmental Impacts (Duration)	2
Unexpected Utility Encounter (Cost)	4
Unexpected Utility Encounter (Duration)	4