ISSN 1392-3730 / eISSN 1822-3605

JOURNAL of CIVIL ENGINEERING and MANAGEMENT

2025
Volume 31
Issue 7
Pages 687–709
https://doi.org/10.3846/jcem.2025.23472

PREDICTION OF SEWAGE PIPELINE CONSTRUCTION DURATION BY INTRODUCING MACHINE LEARNING AND DEEP LEARNING APPROACHES

Sang-Jun PARK[1], Norhane NOUR[1], Kang Young LEE[2], Ju-Hyung KIM[1]✉

[1] *Department of Architectural Engineering, Hanyang University, Seoul, South Korea*
[2] *SAMAN Engineering, Waterworks & Sewage Division 1, Seoul, South Korea*

Article History:
- received 24 October 2023
- accepted 27 September 2024

Abstract. Establishing project costs in construction is crucial for project success, typically done through regression methods for prediction. While these methods are common, novel regression methods are less practiced in construction management. This study explores both traditional and modern regression techniques, analyzing data from 83 sewage pipeline projects in South Korea. The study implemented state-of-the-art frameworks, including hyperparameter optimization and k-fold cross-validation, to evaluate statistic, machine learning and deep learning based regression models using R^2 score, RMSE, MAE, and MSE. Results revealed that performance metrics don't always align with predictive accuracy. For instance, the random forest regressor achieved the best R^2 score of 0.847 but ranked fifth in prediction accuracy. Moreover, polynomial regression outperformed novel methods with a 98.790% accuracy across the validation dataset.

Keywords: construction management, sewage pipeline construction, statistical regression, machine learning regression, deep learning regression.

✉Corresponding author. E-mail: *kcr97jhk@hanyang.ac.kr*

1. Introduction

Civil infrastructure is an indispensable element of all built urban environments as it enables a wide range of human activities and provide public services such as transportation, water supply, sewage, gas, electricity, and power (Doyle & Havlick, 2009). Out of the numerous civil infrastructures providing services to the residents, sewage system is an essential service for modern living that can impact the environment significantly. Sewage pipelines are considered one of the most crucial components of an urban infrastructure system as they preserve public health by draining wastewater from densely populated areas to necessary treatment plants (Malek Mohammadi et al., 2019; Obradović, 2017; Opila, 2011). It is considered as a large infrastructure typically constructed beneath roadways as shown in Figure 1. Consequently, any damage that may occur to pipelines, such as pipe breakage or deterioration, could likely cause damage to roads (Obradović et al., 2023). In Seoul, between 2016 to 2021, a total of 1,431 cases of sinkholes and roads collapsing were reported. Out of the reported cases, 782 cases, were caused by damage or aging of water and sewage pipes which accounted for 54.7% of the total road damage cases (Kim, 2022). Moreover, according to the Ministry of Land, Infrastructure and Transport of South Korea, a total of 8,424 km of water and sewage pipes were installed for more than 40 years ago while 26,350 km were installed for 30 to 40 years. Therefore, there is a need to repair and replace these aged sewage pipes. In 2018, Urban Infrastructure Headquarters announced a management plan for old sewage pipes of 5,000 km by 2021 including a plan for strengthening 73% of the old sewage pipes older than 20 years (Kim, 2022). Moreover, according to the Ministry of Environment Domestic Sewage Division (2021), a plan in motion was announced in 2020 to plan for a large-scale construction project, up to 33,861,387 m of new pipes nationwide, which adds to the 163,098,677 m of existing sewage pipelines.

From the perspective of project managers for sewage pipeline construction projects, risks, such as cost overruns, relating to schedule delays often lead to poor project performance. In general construction projects, post-evaluation reports from 672 completed construction projects revealed that 71% of the projects with delays had caused

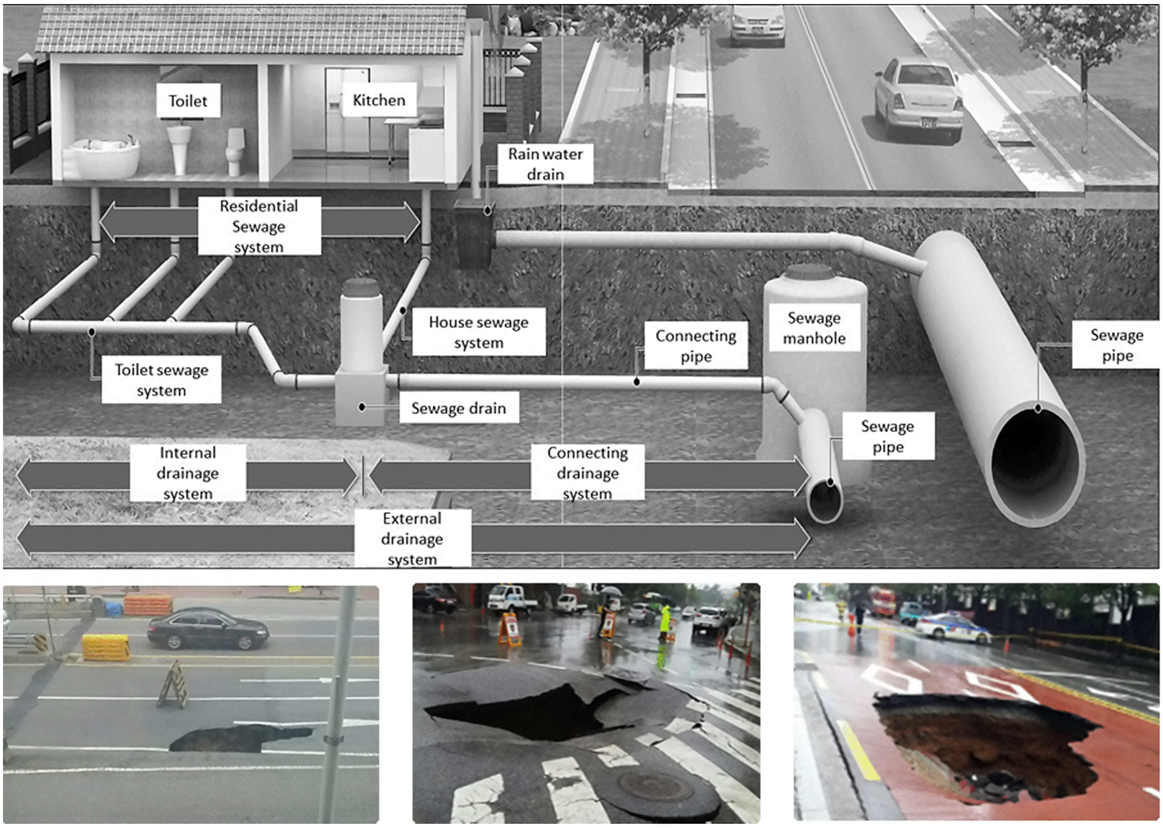


Figure 1. Damage caused to sewage pipelines due to road damages

cost overruns. Schedule delays are a known issue in the field of construction that causes further problems such as: claims, public disruptions and disputes due to road unavailability (Baloyi & Bekker, 2011; Zakaria et al., 2012). For sewage pipeline construction projects, such delays might pose additional concerns related to public health, citizens' well-being and environmental problems due to their importance in preserving public health and protecting the environment. Moreover, a success of a construction project is typically determined during the planning stage where a project manager would make appropriate estimation of general cost of the project (Ganiyu & Zubairu, 2010). A typical cost related project flow is for the project manager to produce a cost estimate consisting of performance, schedule maintenance and budget that creates a standard for determining a success or failure of a construction project. Therefore, in accordance with the upcoming aforementioned sewage pipeline construction projects and to avoid issues related to schedule delays, achieving accuracy in predicting construction duration is essential for progressing toward a successful project completion.

During the planning stage of any construction project, prediction is a vital process that sets the standard for various project related variables that determines project life cycle such as: cost and duration (Tayefeh Hashemi et al., 2020). It is considered a vital process as effectively and reliably predicting the project related variables sets a standard that is used one of the key identifiers for project success or failure (Munns & Bjeirmi, 1996). Prior to the development of artificial intelligent technologies, statistical methods have been used in estimating construction duration and cost that was considered advantageous due to their simplicity in implementations. However, they rely on numerous statistical assumptions (Ghimire et al., 2012; Khedr et al., 2021; Maclin & Opitz, 1999) that hinders accuracy. To increase their reliability, typical statistical approaches demand detailed information relative to a certain project. Such details might not be always available or known in the planning stages of the project when the prediction of the construction duration is mostly needed (Darko et al., 2023). Hence, the need for advanced approaches and more complex models to overcome the uncertainty that usually accompanies the usage of the regression models have been performed and has shown its effectiveness in various scenarios (Mahmoodzadeh et al., 2019; Yuan et al., 2019).

Many previous methodological and technical studies regarding construction duration prediction aims to improve accuracy, efficiency and reliability (Abu Hammad et al., 2010; Kim et al., 2019; Lin et al., 2011; Mahmoodzadeh et al., 2022a, 2022b; Peiman et al., 2025; Pesko et al., 2017; Yeom et al., 2018). However, these studies typically focus on a single discipline, whether that be based on statistics or AI, for their study. For statistical processes, Abu Hammad et al. (2010) used a probabilistic model and statistical regression model based on previous project to predict construction duration of public buildings. A probabilistic approach has been explored in previous studies such as Markov chain method and Monte-Carlo simulation. The probabilistic models have been used to forecast both the ground conditions along the tunnel route and the associated tunnel construction time and costs (Mahmoodzadeh & Zare, 2016). Monte Carlo simulation was used to predict construction cost and time in tunnel construction as well (Moret & Einstein, 2016). For sewage pipeline construction, statistical regression methods had been applied to make an early prediction on the construction cost (Sueri & Erdal, 2022).

The ability of machine learning (ML) and deep learning (DL) models has been proven effective in various construction related problems. Unlike the statistical regression methods, DL and machine learning models improve the process of prediction and are able to overcome the lack of data needed to make accurate estimations (Lee et al., 2016; Akinosho et al., 2020; Saeidlou & Ghadiminia, 2024). Mahmoodzadeh et al. (2022a) implemented Gaussian process regression (GPR) technique to predict construction for digging tunnels in mountainous areas. Prediction of construction duration and cost for green buildings sector was performed using machine learning models such as: deep neural networks (DNN) and support vector regression (SVR) (Darko et al., 2023). Construction cost, which is directly related to construction duration, was estimated for road construction using the least absolute shrinkage and selection operator, K-nearest neighbors (KNN), and random forest (RF) (Abed et al., 2022). Zhang and Li (2024) have made comparison of various machine learning regression methods method previously to predict construction duration of mixed-use buildings with that varied in scale.

In summary, previous studies show that regression model developed from data based on previous projects is a generalized method that has made reliable predictions in the early stages of a project where vital information is scarce and difficult to obtain. Prior to the development of artificial intelligence, statistical based regression method was widely used along with probabilistic methods such as Markov chain and Monte Carlo simulations. Coinciding with the development of technologies, recent studies show an exploration of machine learning based regression methods, such as: SVR, KNN, GPR and RF, made the same predictions. Moreover, DL methods, such as DNN, have also been explored. However, studies have shown that comparison of these regression methods of different topics have been exclusive to a single discipline. Moreover, statistics, ML and DL methods have been widely used in the building construction sector where it was applied to predict cost and duration for various types of buildings with different purposes. Although there were studies that have applied regression methods in the civil infrastructure sector, studies were found to be limited where only a single study concerned sewage pipeline construction with focusing on construction cost.

It is well known that there is a common objective for both numerical, ML and DL based regressions are similar in that it attempts to either predict or set a benchmark. Previous studies have shown that these methods are viable in producing an estimate on scheduling or costs. Within the previous studies, regression has derived initially from statistics gradually being replaced by modern methods. This paper contributes to the body of knowledge by bridging the gap between past and present techniques for performing regression in making estimations and predictions. This research gap emerged from previous studies that have performed predictions using regression; and compared various approaches exclusively from the same discipline. To the best of the authors' knowledge, no studies were found that compared the prediction method from the three different disciplines simultaneously that concerns the same key variables which consequently determines the project success. Moreover, while there was a study that made prediction based on construction cost for other civil infrastructure, limited studies were found that focused on duration for sewage pipeline construction.

In this study, statistical methods used are linear and polynomial regression. ML techniques used in this paper are SVR and RF. Finally, DL techniques used are DNN model and long short-term memory (LSTM). In this study, 83 data acquired from previous sewage pipe construction was deemed sufficient. Dependent and independent variables are identified from the collected dataset that includes: pipe length, construction cost, construction duration where a correlation analysis is performed in order to identify the most influential variable. Traditional and modern methods of producing a regression model is performed where the model is scored according statistical evaluation criteria that are: Pearson correlation (R^2), mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE).

The rest of the paper is organized as follows: Section 2 examines the previous literature divided into two aspects of technical and theoretical discussions, while Section 3 will represent the methodology adopted to carry on the research, explaining the different models that were used to predict the duration of the pipeline construction. In Section 4, prediction results will be presented and analyzed. Finally, the discussion and conclusion will be described in Sections 5 and 6, respectively.

2. Literature review

2.1. Methodological framework from other fields of study

Aside from civil construction and infrastructure, investigating the methodology for performing estimation in other fields provided crucial insight that assisted in improving the general methods used in general civil infrastructure construction. Moreover, by doing so, it provides a good indicator on methods that could potentially be used for planning stage of construction projects.

From the field of mechanical engineering, a study has applied machine learning based regression method for predicting permeability of gas reservoirs (Kamali et al., 2022). In this study, a single method for statistic based and two methods from machine learning was conducted to produce a regression model that shows permeability of gas reservoirs. This study highlights the contribution from this paper in that it is necessary to explore various prediction methods in the field of construction management.

In order to improve the initial methodological framework from previous studies within the topic of construction management, other fields of study were also reviewed. In the field of construction materials, mixture design using alternative fillers for concrete for micro surfacing was performed using machine learning techniques (Gujar & Vakharia, 2019). The study implemented machine learning prediction model to predict fillers composition based on mechanical features of the concrete admixture. The study has highlighted the details of performing the machine learning regression where the hyperparameters were explained in detail. Vakharia and Gujar (2019) conducted a study for predicting the mechanical properties of a high-performance concrete based on the mixture recipe for a Portland cement admixture. While the details of the machine learning regression methods were not mentioned, this study highlighted the importance of conducting the experiment without bias. This was made possible through the use of K-fold cross validation technique and the paper explains in detail the significance eliminating bias from this technique.

Previous studies on relative topics of construction estimation or prediction have found that the general method for performing regression is similar in that the initial dataset is split for training, testing and validation purposes where the split dataset have no influence to the other. Deriving from the reviewed previous studies, Figure 2 shows a conceptual framework for this study. Initial processing is a necessary process to normalise the data. In this stage, K-fold cross validation is applied to split the dataset into train, test and validation dataset to reduce potential bias. The dataset is then applied to the respective regression models deriving from their disciplines.

It was found that a derivative technique from SVR is a regularly used method for performing machine learning based regression along with RF. In particular, hyperparameter optimization was used to tune the machine learning models specific toward the application. Across the studies, it was found that common evaluation methods, such as: MAE, RMSE and Pearson's correlation coefficient, were applied to assess the performance of the applied model. In general, the application for machine learning and DL regression methods followed a similar framework across the various studies. However, whilst the methodologies were applied between the state-of-the-art methods there lacks a study that ties the new methods with the old.

The uniqueness of this study comes in two folds: first, the recently developed methods are performed using the methods from recent studies where the results are compared with the traditional method prior to the development of machine learning. Second, the scarcity of studies relating to sewage pipeline construction, highlighted in the introduction, is performed to assess the feasibility of using reliable methods for the prediction of construction duration using limited available resources. By doing so, this study contributes to the body of knowledge by bridging the gap between the primary sources by which the regression based estimation/prediction method has derived from. Thereby further accenting that: while traditional methods are capable of performing initial estimation there are other methods that provides different results from using recently developed methods.

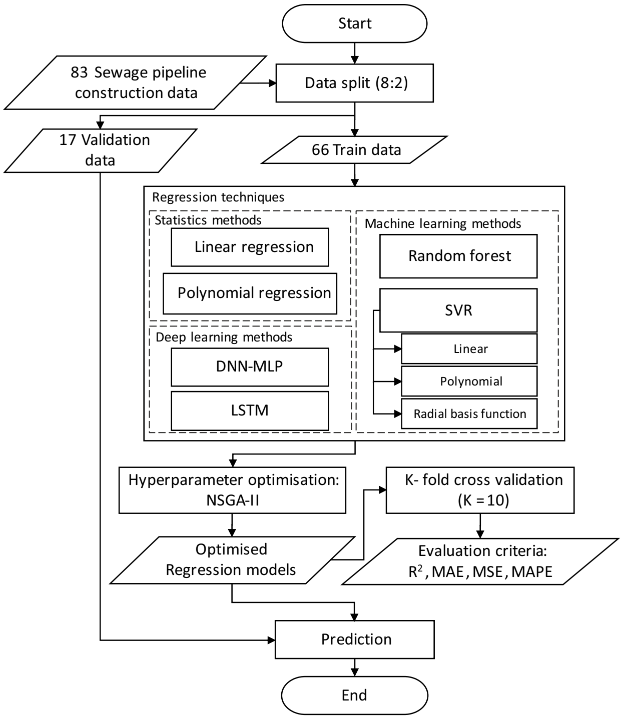


Figure 2. Conceptual flow chart

2.2. Statistical regression methods

2.2.1. Simple linear regression

Simple linear regression is a type of statistical method for modeling the relationship between two variables: a dependent variable and an independent variable, using a linear pattern to the degree of one. The main idea behind linear regression is to fit a straight line to the data points in such a way that it minimizes the difference between the actual data points and the predicted values from the line, by using an error term (Yan & Su, 2009). The linear regression model with one independent variable is represented by the following equation:

$$y = \beta_0 + \beta_1 . x + \varepsilon, \tag{1}$$

where y is the dependent variable and x is the independent variable. β_0 is the intercept and β_1 is the slope, β_0 and β_1 are also called regression coefficient. ε is the error term, accounting to the difference between the observed y and the predicted y.

2.2.2. Polynomial regression

Statistical polynomial regression is a type of regression analysis used to model the relationship between a dependent variable and an independent variable using a polynomial function. Unlike simple linear regression, which assumes a linear relationship, polynomial regression allows for more complex and non-linear relationships between the variables (Rawlings et al., 2001). The polynomial regression model includes using polynomial terms along with linear terms, creating a polynomial equation of a specified degree, which determines the complexity of the model. The general equation to the degree of *K* is represented as follows:

$$y = \beta_0 + \beta_1.x + \beta_2.x^2 + ... + \beta_k.x^k + \varepsilon, \tag{2}$$

where, similarly to the simple linear regression, y is the dependent variable and x is the independent variable. β_0 is the intercept and β_1 is the slope. ε is the error term.

2.3. Artificial intelligent methods

2.3.1. Machine learning in prediction

Machine learning is an artificial intelligence technique that has the ability to train computers to learn from the data, even with smaller datasets (Taye, 2023; MathWorks, 2023). Consequently, ML can train computers to learn from the available data from previously completed projects. ML can, as well, identify the trends and the patterns that exist within the data. Such trends are sometimes overlooked or ignored by simple prediction models (DataFlair, 2022).

There are several ML models, for instance, Mahmoodzadeh et al. (2021) examined machine learning models Decision Tree (DT), GPR, and SVR to reduce the geological uncertainty in tunnels construction durations and costs, using data from previously constructed tunnels and under-studying tunnel observation to train and test the models, depending on only one feature which is the rock mass rating, due to data unavailability. Results proved GPR to be more accurate than SVR, while SVR was more accurate than DT. Rafiei and Adeli (2018) also developed an innovative model based on a back-propagation neural network (BPNN) and Support Vector Machine (SVM) to generate a new concept that estimates the construction cost of low and mid-rise residential buildings, the results of this study showed that costs results estimated by the proposed model were accurate. In another study, aimed at green building projects by Son and Kim (2015), four prediction models were proposed: BPNN, DT, logistic regression (LR), and SVM, to predict the cost and schedule performance of green building projects based. SVM model showed superiority in the accuracy of its prediction over the other three proposed models. Pierdzioch and Risse (2020) proposed using multivariate RF to Forecast the returns of precious metals (gold, silver, palladium, and platinum), where the res. Reports on classical methods, such as Multiple Linear Regression (MLR), indicate that the results based on RF showed higher prediction accuracy than those reported by MLR. Zheng et al. (2023) also examined a RF model optimized by bird swan algorithm to predict the cost of construction projects using cost data from 48 previous construction projects, the results showed that the performance of the proposed optimized RF model was more accurate compared to it to the classical forecasting models such as BPNN, SVM, Stacked Auto-Encoders and Extreme Learning Machine. Meharie and Shaik (2020) examined SVM, NN, and RF models to predict the cost of highway projects in Ethiopia, using data from projects from 2006 to 2018, and then comparing the results using RMSE. The result of this study indicated that the RF model had more accuracy in the prediction with the lowest error value, compared to NN and SVM. Shoar et al. (2022) used MLR, SVR as well as RF to predict engineering service cost overruns in the context of high-rise residential building projects. The results in this context showed that RF has better accuracy followed by SVR and then MLR.

2.3.1.1. Support vector regression (SVR)

SVR is a supervised regression algorithm that is based on the concept of support vector machine, introduced in 1995 (Cortes & Vapnik, 1995). SVM is a supervised machine learning technique used for classification tasks, and SVR is the regression algorithm of the SVM. The main principle of SVM is finding a hyperplane in a high-dimensional feature space that differentiates the data points of different classes. With this perspective, the main objective of SVR is to find the best hyperplane that fits the training set, focusing on minimizing the error within a specified margin (Idowu & Lam, 2020). In SVR, if the training data set is presented as $\{x_i, y_i\}_{i=1}^{l}$, where $x_i \in R^n$ represents n-dimensional input vector and $y_i \in R$ is the one dimensional output value, and the objective is to build a function y = $f(x)$,this function illustrates that the output y_i is dependent on the input x_i.

This function is expressed as:

$$y = w.\varnothing(x) + b, \tag{3}$$

where $w^T x$ is the vector of weight coefficients and b is the bias term.

The regression problem can be represented by the convex optimization problem equation (Yu et al., 2006) which is presented as follows:

$$\text{Minimize } \left(w, b, \xi_l, \xi_i^*\right) = \frac{1}{2} w^2 + C \sum_{i=1}^{1} \left(\xi_i + \xi_i^*\right),$$

$$\begin{aligned} \text{Subject to: } & y_i - w.\varnothing(x) + (b) \leq \varepsilon + \xi_i^*, \\ & \left(w.\varnothing(x) + b\right) - y_i \leq \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 1, \dots, l, \end{aligned} \tag{4}$$

where variables ξ_i and ξ_i^* represent slack variables which define the maximum and minimum training errors allowed within a certain tolerance for error. C is a positive constant that determines the extent of loss incurred when an error is present.

The alternative representation of the nonlinear SVR can be explained as follows:

$$\text{Minimize } \left(\underline{\alpha}_i, \bar{\alpha}_i\right) = \frac{1}{2}\sum_{i.j=1}^{1}\left(\underline{\alpha}_l - \bar{\alpha}_l\right)\left(\underline{\alpha}_j - \bar{\alpha}_j\right)\varnothing\left(x_i\right).\varnothing\left(x_j\right) + \varepsilon\sum_{i=1}^{1}\left(\underline{\alpha}_i + \bar{\alpha}_i\right) - \sum_{i=1}^{1} y_i\left(\underline{\alpha}_i - \bar{\alpha}_i\right),$$

$$\text{Subject to: } \sum_{i=1}^{1}\left(\underline{\alpha}_i - \bar{\alpha}_i\right) = 0,$$

$$0 \le \underline{\alpha}_i \le C,\ i = 1,2,\ldots,l,$$

$$0 \le \bar{\alpha}_i \le C,\ i = 1,2,\ldots,l. \tag{5}$$

Selecting an appropriate nonlinear function on $\varnothing\left(x_i\right)$ and calculating $\varnothing\left(x_i\right).\varnothing\left(x_j\right)$ in the feature space can be challenging. By employing a kernel function $K(x_i, x_j) = \varnothing\left(x_i\right).\varnothing(x_j)$, it allows for the computation of the input space and the generation of inner products in the feature space, overcoming the challenges associated with determining the feature space directly. Functions that meet the Mercer condition in the feature space can be mathematically shown to be equivalent to inner products, making them suitable kernels. Hence, any functions meeting the Mercer theorem can be employed as kernels. The following equations are the kernels used in this research:

Linear Kernel:

$$K\left(x_i, x_j\right) = x_i.x_j. \tag{6}$$

Polynomial Kernel:

$$K\left(x_i, x_j\right) = \left[\gamma\left(x_l, x_j\right) + c\right]^d. \tag{7}$$

Radial Basis Function Kernel:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma\left|x_i - x_j\right|^2\right). \tag{8}$$

Ultimately, the kernel feature enables the expression of the decision function for nonlinear SVR as an equation, which is expressed as follows:

$$f\left(x_i\right) = \sum_{i=1}^{l}\left(\underline{-\alpha}_k + \bar{\alpha}_k\right)K\left(x_i, x_k\right) + b. \tag{9}$$

2.3.1.2. Random forest (RF)

RF is a type of learning technique employed for both classification and regression tasks, introduced by Breiman (2001). The essence of RF lies in the integration of tree predictors, where each tree's outcome relies on the values of a randomly sampled vector. This vector is sampled independently and follows the same distribution across all the trees present in the forest. As with most machine learning based regression algorithms, its primary function is to classify and perform regression using training and testing data. The advantage from this method arises from the decrease in variance and bias without compromising decision accuracy.

In the training stage, the process involves training multiple binary decision trees based on replacement of randomly selected samples from the original dataset. To select segmentation variables and split points, the approach used in this study employs an exhaustive method that involves traversing each feature and all its values. By doing so, the best segmentation variable and split point can be identified, aiming for minimum variance. The quality of the segmentation variable and split point is typically evaluated using the impurity function of the node after segmentation, which represents the weighted sum of impurity for each child node G, which is calculated by the following equations, where equation 12 is a result of substituting Eqn (11) into Eqn (10). Equation (13) is used in the training process of a node within a decision tree endeavors to discover the minimum impurity for each child node G by selecting the most suitable segmentation variable and split point:

$$G\left(x_i, v_{ij}\right) = \frac{n_{left}}{N_s}H\left(X_{left}\right) + \frac{n_{right}}{N_s}H\left(X_{right}\right); \tag{10}$$

$$H\left(X\right) = \frac{1}{N_m}\sum_{i \in N_m}\left(y - \bar{y}_m\right)^2; \tag{5}$$

$$G\left(x, v\right) = \frac{1}{N_s}\left(\sum_{yi \in X_{left}}\left(y_i - \bar{y}_{left}\right)^2 + \sum_{yi \in X_{right}}\left(y_i - \bar{y}_{right}\right)^2;\right. \tag{6}$$

$$\left(x^*, y^*\right) = \text{argmin}_{x,v}\, G\left(x_i, v_{ij}\right), \tag{7}$$

where x_i is the segmentation variable and v_{ij} is the cut value of the segmentation variable; n is the number of training samples of the left and right child node; N_s is the number of all training samples of the current node after segmentation; X the training sample set of the left and right child node; $H(X)$ is the impurity function to measure the impurity of node, where MSE is used in equation 11 as the impurity function.

The Random Forest algorithm operates through parallel integration. Each decision tree functions independently of the others, and the ultimate prediction is derived by averaging the outcomes of all individual decision trees. The final result is calculated using the following equation:

$$f\left(x\right) = \frac{1}{M}\sum_{m=1}^{M} f_m\left(x\right), \tag{8}$$

where $f(x)$ is the final result and $f_m(x)$ is the prediction result from the decision trees.

Out-of-bag (OOB) error estimation is conducted progressively as the forest of regression trees is built. This estimation uses unselected data records, the OOB subset, to test each k tree once it is trained during the bagging process. The OOB subset offers a continuous, unbiased estimate of the general prediction error before validating the accuracy of the aggregated results with an independent testing subset. Additionally, the aggregated results allow for the assessment of each input variable's relative importance in predicting the dependent variable (Barjouei et al., 2021).

2.3.2. Deep learning in prediction

DL is a subset of machine learning that uses deep neural network with many layers. DL is able to automatically extract patterns and dependencies from data, hence its ability to solve complex tasks. In some domains such as natural language processing and bioinformatics among other domains, DL techniques showed a higher performance when compared to ML techniques (Alzubaidi et al., 2021; LeCun et al., 2015).

DL models have been used in various domains. Yu and Yan (2020) examined a DNN-based prediction model. This model was designed based on the phase space reconstruction method and LSTM network for DL and used to predict the stock price. The results were compared as well with the corresponding results from the auto-regressive integrated moving average model, SVR, deep multiple layer perceptron (MLP), and LSTM without the PSR, and they demonstrated that the DNN-based prediction model shows higher estimation capabilities than the other models in stock price prediction. In another study based in Turkey, Bayram et al. (2015) examined RBF and MLP for construction cost prediction, using data obtained from 232 public construction projects. The results of the RBF neural network model showed more accuracy; however, the results predicted by MLP were close to those of RBF. In another study, Mahmoodzadeh et al. (2022d) used optimized LSTM to predict the degree of tunnel wall convergence and compared the results to those of recurrent neural networks (RNN). The LSTM model was proved to be better than other RNN networks at predicting the connections between inputs and TWC. In a study targeting the prediction of engineering cost indexes, Dong et al. (2020) used LSTM neural network. Results showed that LSTM NN had very low prediction errors when compared with the results generated from SVM model. Alshboul et al. (2022) proposed using extreme gradient boosting (XGBOOST), RF, and DNN to allocate green building costs, and it was evident that XGBOOST and DNN outperformed RF with more accuracy. Dang-Trinh et al. (2023) studied SVM, ANN, generalized linear regression, classification and regression-based techniques, exhaustive chi-squared automatic interaction detection as well as DNN to estimate preliminary factory construction costs in Southern Vietnam. DNN was revealed to have the best accuracy among all the examined models. Mahmoodzadeh et al. (2022c) introduced LSTM model optimized by grey wolf optimization algorithms (LSTM-GWO) for tunnel boring machine penetration rate prediction, using data from an Iranian tunnel project, and compared its results with GPR, KNN, SVR, and DT. The results showed that LSTM-GWO achieved the highest accuracy.

The literature highlighted the effectiveness of various machine learning models such as: DT, RF, GPR, SVR, and SVM and DL models such as: LSTM and DNN in various construction-related predictions as well as in different domains. It was evident, as well, that researchers in the construction field have been exploring the application of machine learning techniques to enhance the accuracy of cost predictions as well as duration predictions (Makridakis et al., 2018). That might be due to the fact that the application of ML addresses the challenges related to limited data available, especially during the initial phase of the project (Lee et al., 2016). However, other studies stated that using DL models such as DNN has better accuracy in the field of construction cost estimation which is highly related to duration prediction, due to their ability to learn from past data, to easily analyze the correlation between complex variables and structures. Although, DNN has the ability to replicate the human brain, and was successfully used in various domains; they are frequently used in the prediction of construction duration and cost predictions (Wang et al., 2022).

2.3.2.1. Deep neural network – multilayer perceptron (DNN-MLP)

Deep neural network is a type of artificial neural network (ANN) that is composed of multiple layers. Compared to other neural network structures, the primary advantage of DNN lies in its remarkable nonlinear processing capability (Rumelhart et al., 1988). Its compact and efficient structure for nonlinear mapping enables it to effectively handle mathematical and physical problems that involve larger datasets and more intricate features. Additionally, DNN can fully exploit its multiple hidden layers to train extensive amounts of data, leading to generally higher accuracy in prediction results. The presence of more layers in DNN signifies a more complex model, which exhibits superior nonlinear characteristics and the ability to capture richer features. In theory, the connections between the layers in the network structure are fully linked, allowing neurons within each layer to establish connections with one another. There are several models of DNN, the most widely used model of DNN is multi-layer perceptrons (MLP) (Popescu et al., 2009).

MLPs are commonly used due to their flexibility to fit a wide range of non-linear functions with high accuracy levels. MLPs are feed forward neural networks which are typically composed of several layers of nodes with unidirectional connections, often trained by back propagation. The basic structure of MLP is based on the logic of the biological neuron model. The architecture of MLP includes multiple layers: input layers, several hidden layers, and output layers. Each layer is connected to the next layer, and each layer provides the following layer with the result achieved. For MLP, the dataset was presented as $D=\left\{\left(x_{i,}, y_{i,}\right)\right\}^{n} i=1$, $x_i \in R^{m*l}$, $y_i \in R$, and n is the number of samples. x_i (i = 1, 2, 3, …, n) is the m-dimensional phased feature vector, and y_i is the label of fault. The weight input of j node in the hidden layer can be expressed as h_j, the output of the j node in the hidden later is H_j, and the input of the output layer k note from hidden layers is O_k, these are expressed in Eqns (15), (16), and (17), respectively (Fang et al., 2019):

$$h_j = \sum_{i=1}^{m} w_{ij} * I_i + b_j ; \qquad (9)$$

$$H_j = \tanh\left(h_j\right); \tag{10}$$

$$O_K = \sum_{j=1}^{j} w_{jk} * H_J + b_K, \tag{11}$$

where W_{ij} is the connection weight from the input layer i node to the hidden layer j, I_i (i = 1, 2, 3, ..., m) is the input of MLP, b_j is biased for the corresponding node. The output layer contains K notes (k = 1, 2, ..., K) and the output O_K of the k node in the output layer corresponding to different activation functions.

Moreover, to facilitate the process of training the network and finding the optimal set of weights and biases that reduce the loss function for the given task, an optimization algorithm was used. In this research, Adam optimizer was used. Adam optimizer dynamically modifies the learning rate for each parameter (Kingma & Ba, 2014). These modification and updates of the parameters are expressed by the following equations:

$$m_t = \beta_1 m_{t-1} + \left(1-\beta_1\right) g_1; \tag{12}$$

$$v_t = \beta_2 v_{t-1} + \left(1-\beta_2\right) g_t^2; \tag{13}$$

$$m_t^{\wedge} = \left(\frac{m_t}{1-\beta_1^t}\right) v_1^{\wedge} = \left(\frac{v_t}{1-\beta_2^t}\right); \tag{14}$$

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{v_1^{\wedge} + \in}} m_1^{\wedge}, \tag{15}$$

where β_1 and β_2 are exponential decay rate, with values of 0.9 and 0.999 respectively. The correction biases for m_t and v_t are $m_1^{\wedge}$ and $v_1^{\wedge}$ respectively.

2.3.2.2. Long shot-term memory (LSTM)

LSTM are a type of recurrent neural network (RNN). RNN is a type of neural network that understands and learns from the context of sequential data; however, RNN faces some challenges such as vanishing or exploding gradients, which can hinder the learning process, as well as, limited memory capacity, making it difficult to capture long-term dependencies in sequences. LSTM was first proposed by Hochreiter and Schmidhuber (1997), designed to address the challenges of processing sequential data of the traditional RNN by incorporating a specialized memory cell and gating mechanisms. The architecture of LSTM consists of: an input layer, an output layer, and hidden layers. The input data passes through the input layer and then goes to the hidden layers. The hidden layer captures the information and processes it over time, and only useful information is retained, and then output layer The hidden layer is the most complex layer as it consists of several gates and a memory state unit (Marino et al., 2016).

Forget gate: The information passes through the forget gate and it controls which information will be discarded from the previous layer and which information will be retained; and then sent it to the input gate. The equation of the forget gate is represented below:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right). \tag{16}$$

Input gate: In this step, the input gate determines the information that will be remembered from the new information, and then updates it to be stored later in the memory cells. The equation of the input gate is as follows:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right). \tag{17}$$

Output gate: In this step, the output gate determines the output of the model and the contribution of the control unit state (C_t) to the hidden layer elements. It begins by using the sigmoid activation function to calculate the initial output. This output is then transformed to a range of –1 to 1 through the application of the *tanh* function. Finally, it is multiplied with the output of the sigmoid function to obtain the final result. This can be expressed as follows:

$$O_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right); \tag{18}$$

$$h_t = O_t \cdot \tanh\left(C_t\right). \tag{19}$$

The memory cell: The Memory Cell is positioned at the top and employs the tanh function to create fresh candidate values. It integrates the input information from the Input Gate with the existing state information in order to modify the memory state. Its role is to determine the presently stored information and the information that will be passed on to the next stage. By utilizing historical data, it is capable of making predictions about future data. The memory cell is represented by the following equation:

$$\widetilde{C_t} = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right). \tag{20}$$

In the above-mentioned equations, f_t, i_t, O_t and $\widetilde{C_t}$ are the forget gate, the input gate and the output gate respectively. σ represents the sigmoid activation function and h_{t-1} is the output of LSTM at timestep. t – 1, x_t is the input data; while W_f, W_i, W_o and W_c are the weights of the forget gate, the input gate, the output gate and the memory cell respectively. C_t is the output intermediate cell and b represents the bias for each gate.

While the literature offered valuable insight into the application of ML and DL in prediction, the findings of the literature contradict each other regarding which prediction model is the most accurate in the estimation of the construction duration generally. In addition, research on the field of the sewage pipeline has been scarce from the start. Due to this gap, it is important to conduct research concerning sewage pipeline construction duration prediction, exploring the potential of the application of ML and DL, for the aforementioned purpose, will be explored. After examining the literature, for more accurate prediction, this study will employ advanced machine learning and DL models for more accurate estimations, namely: SVR, RF, LSTM, and DNN, as these models were proven to be effective in related construction prediction tasks.

2.4. Statistical evaluation criteria

Statistical evaluation criteria are used after each method of prediction in order to assess the performance and evaluate the accuracy of the method. Using these criteria helps determine the reliability of the forecasting models and confirm the fittest method for the data (Behnia et al., 2013; Kamali et al., 2022). In this research, the criteria used were Pearson's correlation (R^2), which is also known as the coefficient of determination, MSE, RMSE and MAE, the following equations represent each criterion, respectively:

$$R^2 = \left(\frac{\sum_{i=1}^{n} (f(\left(x_i\right) - \bar{f}\left(x\right))\left(f^*\left(x_i\right) - f^*\left(x\right)\right)}{\sqrt{\sum_{i=1}^{n} (f\left(x_i\right) - \bar{f}\left(x\right))^2 \sum_{i=1}^{n} (f^*\left(x_i\right) - \bar{f}\left(x\right))^2}} \right)^2; \quad (27)$$

$$MSE = \left(\frac{1}{n}\right) \sum_{1}^{n} \left(f\left(x_i\right) - f^*\left(x_i\right)\right)^2; \quad (28)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\widehat{y_i} - y_i\right)^2}{n}}; \quad (29)$$

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \left|f\left(x_i\right) - f^*\left(x_i\right)\right|, \quad (30)$$

where $f(x)$ is the actual value and $f^*(x)$ is the forecasted value. $\bar{f}(x)$ are the means of actual and predicted and n is the number of datasets.

2.5. Hyperparameter optimization using NSGA-II

In recent machine learning and DL studies, hyperparameter tuning or optimization is a necessary process to obtain the most suitable machine learning and DL models to maximize the performance and results. The parameters to optimize is unique to the regression method mentioned above. Often, machine learning and DL procedures are referred to as a 'black box' problem where fundamental calculations are unknown that are still required to produce the most appropriate types of results for the sake of stakeholders (Karl et al., 2023). On the contrary to the uncontrollable parameters there are parameters that are able to be controlled.

Optimization is a mathematical, and often computational, process of obtaining the best possible solution within a defined set of constraints. It typically involves a form of optimization algorithm that includes control values and the results. For machine learning regression, the control parameters depend on the model such as SVR and RF. For DL, typical control parameters consist of epoch and learning rates. However, tuning these parameters in a conventional method of trial-and-error is time consuming as it needs to be controlled per run. Optimization algorithms, such as evolutionary optimization or Bayesian optimization processes, aims to reach the best suitable results based on the outcome that is the statistical evaluation criteria explained below. As machine and DL models compose of multiple inputs and outputs it can generally be viewed as a multi-objective optimization problem. To that end, in this study, an exploration of hyperparameter optimization is performed by non-dominated sorting genetic algorithm (NSGA-II).

NSGA-II was initially created to address limitation of its predecessor that required high computational complexity, non-elitism approach and a need for specifying a sharing parameter (Deb et al., 2002). To address these issues, it introduced elitism in order to preserve the best solutions per iteration. It also includes a method for maintaining diversity by introducing a crowding distance mechanism on the pareto-optimal solutions during the search process. The primary components are as follows:

- **Fast non-dominated sorting:** The population are sorted into non-dominated fronts where solutions within the same front do not dominate each other.
- **Crowding distance calculation:** Crowding distance computed for individual solutions by measuring closeness of neighbours in objective space.
- **Elitism:** Elitism strategy retains the best solution to guarantee the best genes are passed on to the next generation.
- **Binary tournament selection:** Selection is performed by a pair of solutions based on the rank.
- **Crossover and mutation:** Generate an offspring to introduce new solutions by combining two parent solutions and introduce a small change to offspring solution.
- **Forming the next generation:** Combine parent and offspring populations to select best solution to form the next generation.

The input variables for the regression models used in this study for the optimization is described in Table 1.

Table 1. Regression models and their respective hyperparameters

Model	Variable 1	Variable 2	Variable 3
Linear regression	N/A	N/A	N/A
Polynomial regression	Degree	N/A	N/A
SVR linear	Regularization parameter (C)	Epsilon	N/A
SVR polynomial	Regularization parameter (C)	Epsilon	Degree
SVR RBF	Regularization parameter (C)	Epsilon	Kernel coefficient (Gamma)
Random forest	estimators	Max depth	Min samples split
DNN-MLP	Hidden layers	Neurons	Dropout rate
LSTM	LSTM units	Dropout rate	N/A

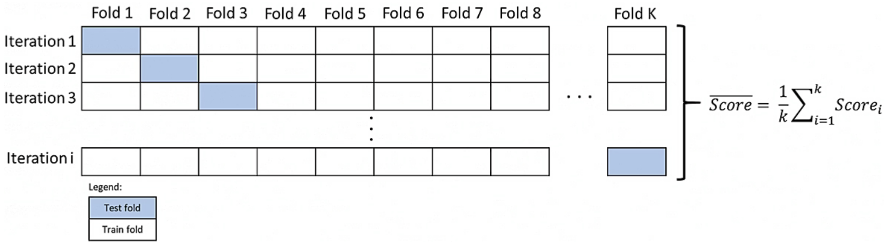


Figure 3. K-fold cross validation visualisation

2.6. K-fold cross validation

Cross validation is a frequently used numerical process that ensures the performance of a machine learning model by keeping it un-biased. A well-known type of cross validation is the K-fold cross validation where a 'fold' is subset of a dataset used only a single time per iteration. In this method, datasets are divided equally where the model is trained using different folds. Validation and training are performed with the '*k*' and '*k* – 1' folds, respectively, per iteration. As mentioned previously, this method ensures to eliminate bias in the resulting regression model and have been shown to be effective in fields of study other than that of construction management (Vakharia & Gujar, 2019). Conveniently, this method is openly available to implement within in the Sci-Kit learn package. A general visualization of K-fold cross-validation process is shown in Figure 3.

The score for each iteration is extracted and calculated to produce an overall performance for the model. Formula for overall score is shown on the right of Figure 3 where: calculated from the formula below where *k* is the number of folds and *i* is the iteration.

3. Methodology

This research aims to find the most accurate prediction method for the construction duration of sewage pipelines. To achieve this aim, a methodology of two stages was developed, Figure 2 shows the approach applied for this research. In stage one, using the data from 83 previous sewage pipeline construction projects from 1999 to 2022, the construction duration of sewage pipelines was predicted. The parameter that was considered in this research was the length of the pipes as it was the most available parameter for all the 83 previous projects.

Predictive analyses were conducted using a range of techniques: statistical regression methods (linear and polynomial) were executed in SPSS 27, machine learning models, and DL models were implemented in Python 3.6 on the Jupyter Notebook platform. Scikit-learn was employed to execute three different SVR variants, including linear, polynomial, and radial basis function kernels, in addition to the RF. For DL tasks, TensorFlow 2.12 was utilized to execute both DNN-MLP and LSTM models. Moreover, to train and test the machine learning and DL models, the data available, represented in Figure 3, were split into 80% of the data for training and 20% of the data for validation, this data distribution is shown in Figure 4. Each of the prediction methods is explained in detail in this section. In stage two, the results were compared and evaluated using four statistical evaluation criteria to assess the performance of each method, in this research, the statistical evaluation criteria used are: R^2, MSE, RMSE and MAE that provides a suffice indicator of the model performances according to previous research (Bui et al., 2020).

3.1. Data acquisition and variable decision

Data is collected from previous construction works. Data that was available for collection consisted of pipe length, total construction cost and construction duration from 83 sewage pipeline construction sites. The data is shown in Table 2. The data collected consists of the physical properties such as the pipe length, minimum and maximum pipe diameters. It also includes the total construction cost, in euros, for the specific project along with the construction dates where the construction duration, in months, was extracted. There are variances in the data where the most expensive construction project does not reflect the scale of the construction nor the duration. However, there are some definite signs where the scale of the construction corresponds to the duration of the construction.

Out of the collected data, the dependent data was chosen as construction duration as this is dependent on the construction scale, i.e., longer the pipe length or larger the construction cost the longer the duration. From the independent variables that are: sewage pipe length and construction cost, correlation analysis was performed in order to identify the more influencing factor. The correlation analysis showed that pipe length and the construction cost was 0.895 and 0.810, respectively. This showed that pipe length was more influential to the dependent variable over the construction cost. Therefore, pipe length was chosen to be the main independent variable in this study.

The initial data used is shown in Figure 4a where the acquired data was split to train and validation data by the order of 8:2 that is shown in Figure 4b. The graph shows a normalized data for construction duration and the pipe length, that are the *x* axis and the *y* axis respectively, of the corresponding project. Square data represents the training data and the red triangle represents the validation data used.

Table 2. Sewage pipeline construction project and variable

Project	Minimum pipe diameter (mm)	Maximum pipe diameter (mm)	Pipe length (km)	Total construction cost (EUR)	Construction dates	Construction duration (Months)
Project 1	300	400	28.5	6,295,800	1999.11.23~2002.06.09	30
Project 2	300	1350	3.32	3,150,000	2000.10.21~2003.05.22	19
Project 3	250	600	10.19	1,876,000	2001.07.10~2003.03.30	20
Project 4	200	400	9.34	8,825,600	2001.05.07~2003.03.26	23
Project 5	200	400	21.2	7,210,000	2002.08.14~2004.12.10	28
Project 6	700	900	1.99	3,290,000	1999.5.27~2000.10.06	16
Project 7	200	500	23.6	5,740,000	2003.07.07~2006.04.31	33
Project 8	200	600	18.6	6,160,000	2003.04.02~2005.08.10	28
Project 9	250	1100	10.1	12,300,400	2003.11.27~2005.07.10	20
Project 10	300	1000	8.5	5,810,000	2004.01.07~2005.09.13	20
Project 11	150	1000	20.6	5,460,000	2004.05.17~2006.07.15	26
Project 12	250	1200	27.6	13,160,000	2004.07.22~2007.02.10	30
Project 13	300	1800	45.9	33,950,000	2004.11.01~2009.01.31	50
Project 14	250	700	15.3	8,960,000	2004.12.01~2007.02.01	26
Project 15	250	1000	15.2	9,590,000	2004.12.27~2007.05.20	29
Project 16	80	500	14	5,040,000	2005.01.05~2007.04.18	27
Project 17	250	1000	24.8	7,980,000	2005.03.16~2007.12.20	33
Project 18	300	1200	5.5	3,500,000	2005.06.15~2007.07.09	25
Project 19	150	1200	53.8	14,700,000	2005.06.28~2007.12.04	29
Project 20	150	1000	5.3	7,910,000	2005.11.01~2007.12.07	25
Project 21	200	1000	45.7	21,560,000	2006.03.02~2009.10.03	43
Project 22	200	600	31.7	12,880,000	2006.04.24~2009.03.25	34
Project 23	200	800	23.2	9,730,000	2006.08.28~2009.06.18	33
Project 24	150	1200	7.5	8,491,000	2003.10.10~2005.06.31	20
Project 25	150	400	98.1	22,631,000	2006.02.13~2011.05.19	62
Project 26	200	1100	28	12,110,000	2006.12.04~2009.12.04	36
Project 27	200	600	18.2	5,320,000	2006.08.31~2009.04.30	31
Project 28	200	800	16.1	5,320,000	2007.02.09~2009.05.22	27
Project 29	200	800	19.4	5,740,000	2007.08.13~2010.08.02	35
Project 30	250	1350	19.1	6,860,000	2007.12.20~2010.09.19	32
Project 31	200	1000	45	22,050,000	2007.12.27~2012.01.03	48
Project 32	200	600	29	12,460,000	2008.04.18~2011.03.19	35
Project 33	200	600	26	10,080,000	2008.02.13~2011.02.31	36
Project 34	200	800	42	22,680,000	2008.06.02~2011.05.01	35
Project 35	200	700	9.8	4,179,000	2008.09.22~2010.12.24	27
Project 36	200	400	41.5	17,983,000	2008.05.19~2011.06.02	36
Project 37	300	1200	25.3	40,600,000	2008.11.11~2012.04.01	40
Project 38	200	600	53.5	20,300,000	2008.10.01~2011.07.10	33
Project 39	200	300	41.5	12,670,000	2008.07.29~2011.07.13	36
Project 40	200	1000	8.9	6,045,900	2008.12.01~2010.08.25	20
Project 41	300	1000	10.2	16,100,000	2009.07.30~2012.02.30	30
Project 42	250	1200	18	13,580,000	2009.07.20~2012.10.10	38
Project 43	200	1200	11.9	6,520,500	2009.04.16~2011.06.15	25
Project 44	200	1200	55.2	25,844,000	2009.12.23~2014.01.15	48
Project 45	300	1500	37.2	26,306,000	2009.03.04~2013.09.18	53
Project 46	300	400	14.5	12,817,000	2010.06.16~2013.04.31	33
Project 47	300	1350	29	22,400,000	2010.05.20~2014.03.28	45
Project 48	N/A	N/A	30.6	14,560,000	2011.10.07~2014.07.31	34
Project 49	450	1200	44.2	17,855,600	2011.03.14~2014.08.12	40
Project 50	200	600	20.3	7,210,000	2011.08.31~2013.10.30	25

End of Table 2

Project	Minimum pipe diameter (mm)	Maximum pipe diameter (mm)	Pipe length (km)	Total construction cost (EUR)	Construction dates	Construction duration (Months)
Project 51	80	600	55.6	15,960,000	2011.09.20~2015.12.24	50
Project 52	80	250	39.43	7,147,000	2011.06.07~2014.05.21	36
Project 53	80	250	6.19	5,530,000	2012.05.04~2014.07.09	26
Project 54	N/A	N/A	30.7	18,179,000	2013.06.03~2016.08.18	39
Project 55	300	1500	15	5,390,000	2012.10.05~2014.11.07	25
Project 56	80	1000	32	8,824,200	2013.02.13~2016.10.30	43
Project 57	N/A	N/A	3	1,904,000	2013.06.10~2015.03.13	21
Project 58	200	1200	33.7	14,840,000	2012.04.27~2015.04.30	36
Project 59	80	900	29	18,200,000	2013.05.27~2016.04.28	35
Project 60	200	1200	31	22,680,000	2013.08.12~2017.03.28	42
Project 61	200	1000	24.5	14,140,000	2013.04.15~2016.05.03	36
Project 62	N/A	N/A	88.1	30,590,000	2014.05.19~2019.05.28	60
Project 63	800	1000	19.8	7,175,000	2014.12.12~2017.04.15	28
Project 64	200	400	4.4	5,040,000	2015.09.21~2017.06.30	20
Project 65	150	400	18.2	9,520,000	2015.11.03~2018.01.31	27
Project 66	200	400	97.86	39,550,000	2016.01.25~2018.11.24	60
Project 67	N/A	N/A	18.1	6,790,000	2016.10.10~2019.04.30	30
Project 68	200	250	34.2	15,431,500	2017.07.14~2020.07.31	37
Project 69	N/A	N/A	24.8	8,890,000	2015.08.17~2018.09.31	36
Project 70	200	600	9.7	12,700,800	2017.03.02~2019.11.11	32
Project 71	800	1000	4.65	5,352,900	2017.03.28~2019.01.14	22
Project 72	200	200	13	5,135,200	2017.06.16~2019.12.21	29
Project 73	200	200	19.88	12,600,000	2017.12.11~2020.06.24	30
Project 74	200	600	33.2	10,500,000	2017.12.18~2020.10.04	33
Project 75	150	600	19.28	10,360,000	2018.06.20~2021.01.16	30
Project 76	200	200	24.8	14,910,000	2018.07.11~2021.09.09	38
Project 77	N/A	N/A	42.52	31,150,000	2019.02.18~2022.04.30	38
Project 78	200	1500	43.7	19,320,000	2019.02.18~2020.12.16	36
Project 79	100	250	13.6	13,580,000	2020.12.01~2022.12.30	24
Project 80	N/A	N/A	14.65	10,570,000	2019.11.28~2021.11.16	24
Project 81	N/A	N/A	30.7	13,510,000	2019.08.26~2022.05.25	32
Project 82	N/A	N/A	16.2	13,020,000	2019.12.27~2022.02.13	29
Project 83	N/A	N/A	30.5	22,400,000	2019.05.29~2022.01.27	38

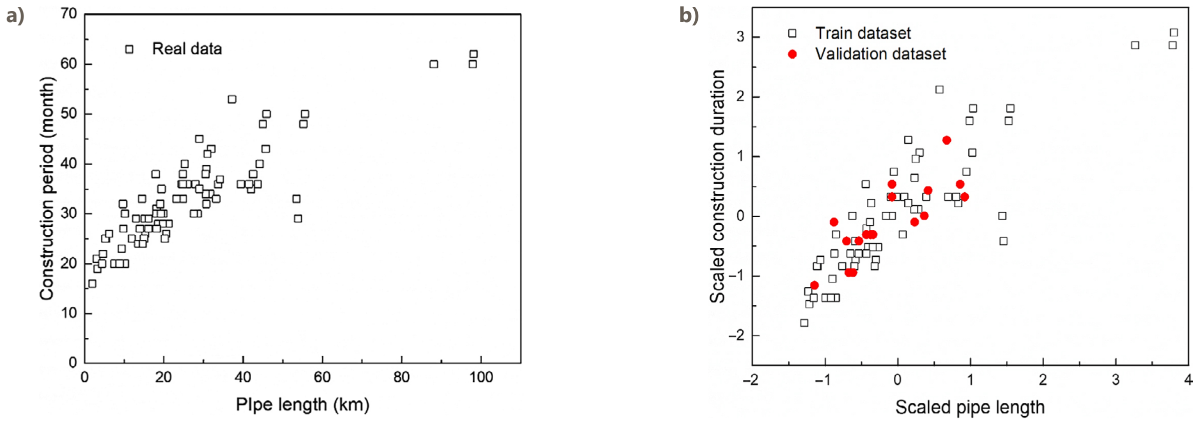


Figure 4. Sewage construction duration data: a – Real 83 sewage construction duration data; b – Distribution of standard scaled training and test data

3.2. Statistical regression methods

As mentioned previously in this paper, statistical methods were conducted using Python within the Jupyter Notebook environment. This was due to the implementation of hyperparameter optimization for the polynomial regression and cross validation. Across all regression models, dependent and independent variables must be chosen. Upon initially analyzing the data, construction duration was chosen as the dependent variable as this is the target variable to be achieved. Pipe length was chosen as the independent variable as it varies according to the project scale. To maintain consistency in assessing regression model performance, the same statistical performance criteria was used consisting of Pearson correlation, MAE, RMSE and MSE.

As mentioned previously, linear regression has no hyperparameter to be optimized and therefore was neglected in the process. Unlike linear regression, polynomial regression has a single parameter, degree, that has potential to be tuned to determine the nature of the regression model.

With these variables R^2 value is calculated to be 0.823 which shows a strong correlation between the pipe length and the construction duration. For the regression model to be deemed significant the comparison between F-test and P-value is used. For the linear regression model, F-test was found to be 187.22 and p-value was calculated at 0.000. As p-value is less than that of F-test, it is confirmed that the regression model is significant. Similarly, for the polynomial model, R^2 value, F-test and p-value is calculated at 0.832, 187.22 and 0.000, respectively. That shows significance in the polynomial regression model produced.

3.3. Artificial intelligence methods

Machine learning and DL methods for prediction sewage pipeline construction duration was performed using Jupyter notebook powered by Python. In this research, Python 3.10 was used along with Scikit-learn 1.5.1 for performing SVR. Within the SVR function, there are variables that determines the resulting regression model such as: degree of freedom, coef0 which is the independent term in the polynomial kernel function, gamma value that is a kernel coefficient, independent term in kernel function and epsilon that specifies the epsilon-tube which no penalty is associated in the training loss function against points predicted from the actual value. An example of the the function entered into the Jupyter notebook is shown in Figure 5. The mentioned control parameters are optimized using NSGA-II that was performed using a package called Optuna 3.6.1. The objective function for '*C*' is performed using the wrapper method initialized uniformly using logarithm function with the boundary from 0.001 to 1000. For epsilon, the same method was used with the boundary from 0.01 to 10. Extra variables need to be taken into consideration for polynomial and RBF kernels such as the degree of freedom and gamma, respectively. Degree of freedom for the polynomial kernel is an interger between 2 and 5. For the RBF kernel, gamma followed the uniformly distributed logarithm function between 0.0001 and 0.1. The aim the optimization function was to minimize the resulting MSE reflecting the accuracy of the generated model. The number of iterations performed for the optimization was chosen as 2000 generations to cover as much variables as possible and is consistent throughout the hyperparameter optimization for all regression models bar linear regression. The objective function and the respective boundaries are shown in Table 3.

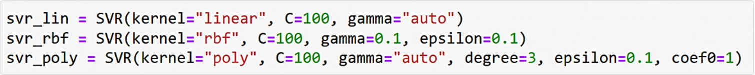
An alternative machine learning regression model investigated in this study is the random forest regression. The same python package, Sci-kit learn 1.5.1, was used to execute the machine learning model. The control parameters for this model are the number of estimators, maximum depth and minimum sample split. Number of estimators are the number for trees in the forest where the values for the optimization was chosen as a range of integers between 50 and 200. Maximum depth is the maximum depth of the trees where the range was chosen as integers between 5 and 30. Minimum sample split is the number of samples required to split an internal node which was chosen as integers between 2 and 10. The summary of the optimization function is shown in Table 4 and the result from the hyperparameter optimization using NSGA-II for SVR linear, polynomial, RBF and random forest are shown in Table 5.

Table 3. Optimization function for the SVR regression function

Variables	Target
Number of generations	500
Output	Mean squared error
Target	Minimise
Regularization parameter	0.001 < C < 1000
Epsilon tube	0.01 < epsilon < 10
Degree (polynomial)	2 < degree < 5
Kernel coefficient (RBF)	0.0001 < gamma < 0.1

Table 4. Optimization function for Random forest regression function

Variables	Target
Number of generations	500
Output	Mean squared error
Target	Minimise
Number of trees	50 < n_estimators < 200
Maximum tree depth	5 < max_depth < 30
Minimum sample split	2 < min_samples_split < 10



```
svr_lin = SVR(kernel="linear", C=100, gamma="auto")
svr_rbf = SVR(kernel="rbf", C=100, gamma=0.1, epsilon=0.1)
svr_poly = SVR(kernel="poly", C=100, gamma="auto", degree=3, epsilon=0.1, coef0=1)
```

Figure 5. Example of SVR function specification within Jupyter notebook

Table 5. Optimised hyperparameters for respective machine learning regression models achieved with NSGA-II

Machine learning model	Variable 1	Variable 2	Variable 3	Best generation
Linear	C = 0.597	Epsilon = 0.489	N/A	469
Polynomial	C = 0.0012	Epsilon = 0.0974	Degree = 3	398
SVR	C = 507.432	Epsilon = 0.532	Gamma = 0.0121	330
Random forest	N estimators = 97	Max depth = 7	Min samples split = 2	357

In addition to the machine learning techniques, DL techniques for producing regression models are also explored in this research. Again, Jupyter notebook is used to code the DNN-MLP and LSTM model using an openly available package called Tensorflow 2.10.1. Tensorflow-keras is used as an import package where layers, sequential and functions for individual hidden layers is used for DNN-MLP and to call the LSTM function for the LSTM model. For both optimizing hyperparameters of DL models, the batch size and epochs are 32 and 100, respectively.

The hyperparameter optimization is kept consistent with the machine learning and statistics regression model where NSGA-II is used with 50 iterations to identify the best performing hyper parameter. The parameters used for the DNN-MLP model are the number of hidden layers, number neurons and dropout rate. As the regression task is relatively simple compared to that of image or language processing, the boundary for the number of hidden layers covers a range of 1 to 3 as an integer. Boundaries for the number of neurons is between 10 and 100 in integers and the dropout rate is between 0 and 0.5 in uniform fashion. For the LSTM model, the parameters controlled are the number of LSTM units and dropout rate with a boundary of 10 to 100 and 0 to 0.5, respectively. The number of LSTM units are generated as an integer and dropout rate is generated uniformly. The boundaries for the optimization function and the optimized hyperparameters are shown in Tables 6 and 7.

In order to analyze the appropriate training and validation for this model, loss, given in mean squared error between training and validation data is compared as shown in Figure 6. Moreover, validation data has no influence on the training part of the modelling. Therefore, the training DNN-MLP model is performed with the training data only where it is validated against the validation data. In Figure 6, towards the end of the epochs, train loss and validation loss are closely converge visually signifying the validity of the DNN-MLP model. Similar to the DNN-MLP model, training for the LSTM model is also visually signified by the close convergence of the loss function towards the end of the epoch. However, for the models the training loss and validation loss never actually quite meet making it plausible for the model to be overfitting.

Analyzing loss function given by mean squared error to the training by the LSTM model is similar to analyzing loss and accuracy for DNN-MLP model. Figure 7 shows the loss function according to the LSTM model where training wasn't as complete as the DNN-MLP model. Again, the training loss and validation loss never meet making it plausible for over or underfitting.

In this section, a brief overview for the statistical analysis, machine learning and DL processes using a commercially available software has been conducted. In order, statistical regression was performed first that included hyperparameter optimization was performed for the polynomial regression only with degree as its optimizing variable. Both regression models have shown validity through the comparison between F-test and p-value deeming the resulting model significant. For machine learning, boundaries of control parameters were highlighted required for the hyperparameter optimization process where the most optimal hyperparameters were found that resulted in the minimum MSE. For deep learning models, the same hyperparameter optimization was performed using NSGA-II across all regression models. The results from the hyperparameter optimization is shown that determined the shape of the deep learning models.

With the optimized hyperparameter, k-fold cross validation is performed to obtain the average score of the process. The scores include: Pearson's correlation value, MSE, MAE and RMSE as the performance criteria to analyze the fitting of the regression models.

Table 6. Optimization function for deep learning regression functions

Deep learning model	Variables	Target
All	Number of generations	500
	Output	Mean squared error
	Target	Minimise
DNN-MLP	Number of hidden layers	1 < n_hidden_layers < 3
	Number of neurons	10 < n_neurons < 100
	Dropout rate	0 < dropout_rate < 0.5
LSTM	Number of LSTM units	10 < n_lstm_units < 100
	Dropout rate	0 < dropout_rate < 0.5

Table 7. Optimised hyperparameters for deep learning regression models achieved with NSGA-II

Deep learning model	Variable 1	Variable 2	Variable 3	Best generation
DNN-MLP	N_hidden_layers = 1	N_neurons = 79	Dropout_rate = 0.402	282
LSTM	N_lstm_units = 99	Dropout_rate = 0.078	N/A	345

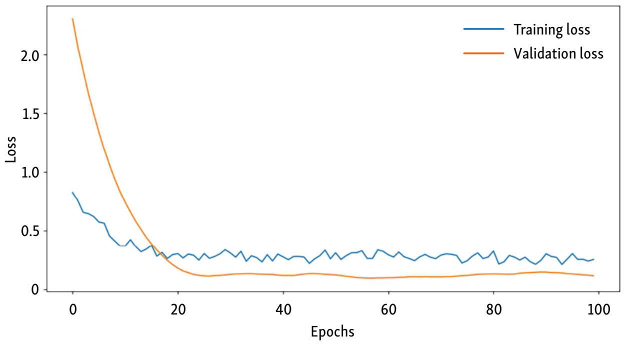


Figure 6. Training and validation loss for DNN-MLP model training

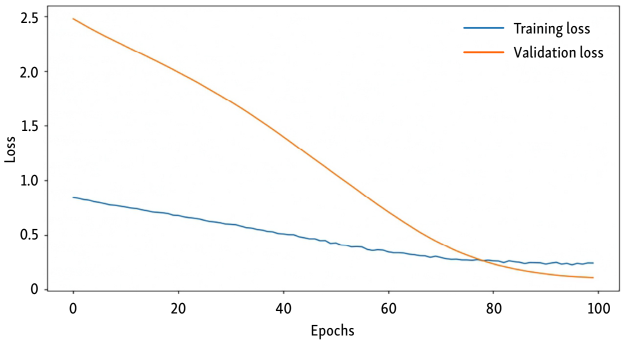


Figure 7. Loss function for training and validating LSTM model

4. Results

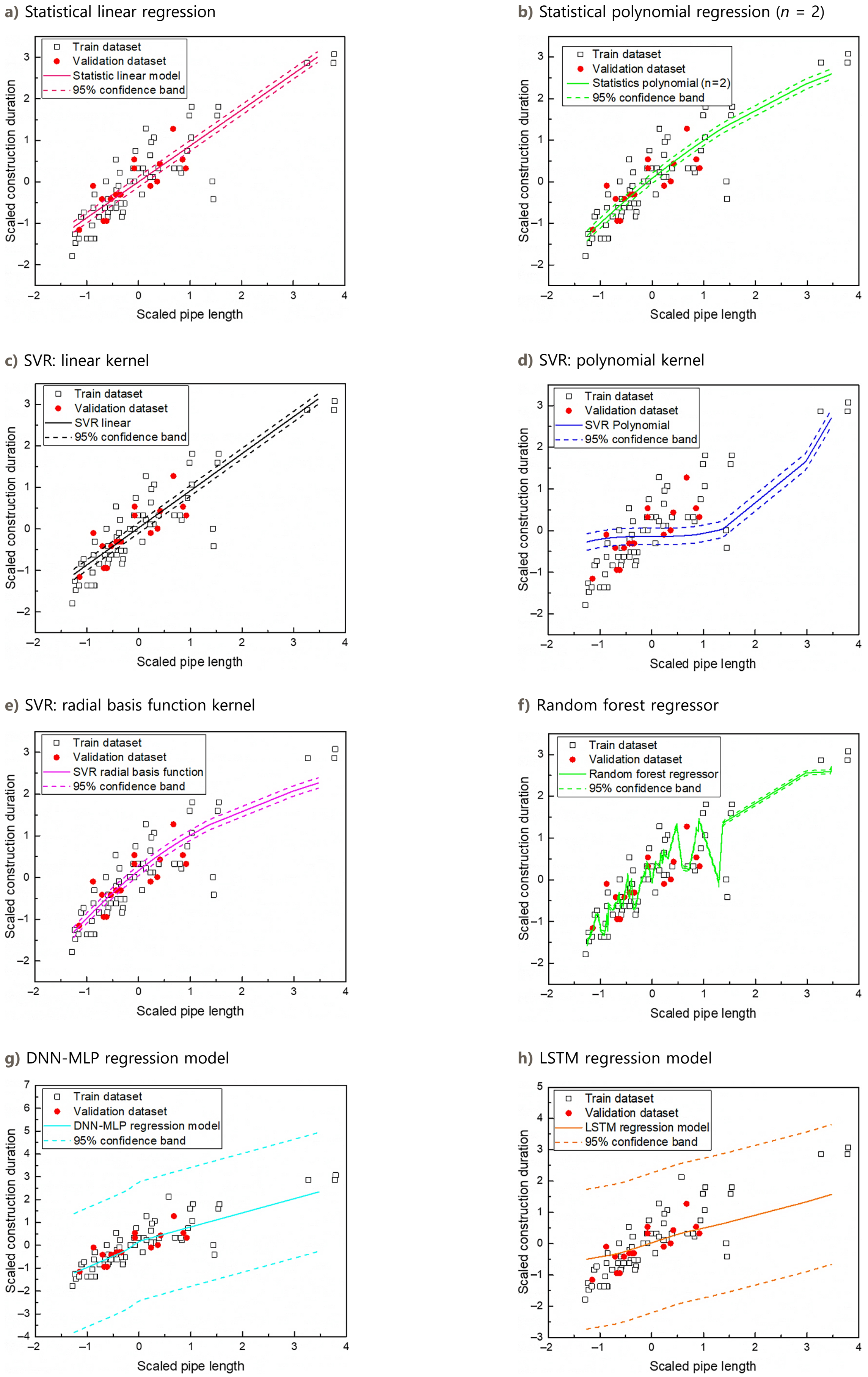
In this paper, predictions were made by using various regression functions with data from 83 previous pipeline construction projects. Where pipe length was chosen as the main independent variable and actual construction duration was chosen as the dependent variable. All regression models were performed with Python with a commercially available packages for performing statistics, ML and DL based regressions. As mentioned previously, the original data was collected and split by 8:2 ratio to obtain training and validation datasets where the latter was strictly removed from the modelling process to ensure there was no influence on producing the regression models. All models underwent the hyperparameter optimization using NSGA-II where the appropriate hyperparameters were chosen based on the lowest MSE after 500 generations. Then, K-fold cross validation was performed in case the result from the traditional train test split function is biased. The average score across the 10-fold cross validation is calculated to achieve the final performance criteria.

For ML, 3 kernels of SVR were performed where linear, polynomial and radial basis function. Furthermore, an openly available regressor function, RF, was also performed. For deep learning, DNN-MLP and LSTM was used against the training dataset where the same dataset was used to predict the construction duration to extract the construction duration that resulted in respective regression models. The hyperparameters for these regression functions were optimized using NSGA-II as explained in the previous section. From the individual functions, respective regression models were created using the training data deriving from an 8:2 split from the original data. Using the regression model, a prediction of the construction month is made from the validation data where an average accuracy percentage was calculated.

The results extracted for respective regression models were organized and represented using OriginPro 2016 as shown in Figure 12. As mentioned previously in the paper, the regression models were strictly performed against the training dataset that included k-fold cross validation to remove bias. Along with the main regression model, 95% confidence interval was included. Moreover, with the resulting regression model, validation dataset was used to perform further statistical analysis. Within the graphs, the squares represent the training data and the red circles represent the validation datasets. X axis shows the scaled pipe length and the y axis shows scaled construction duration. The solid lines represent the main regression model produced by respective methods and the dashed lines represent the 95% confidence interval band. The spread of the training dataset points and validation dataset point around the regression line indicates the fitting of the model against the data. For the statistic and machine learning methods, the confidence band is narrow indicating that the model is confident in making predictions. However, visually, this is not true aside from the DL methods. The validation data points fall outside the confidence band and, therefore, the reliability of the regression models produced is questionable.

Figures 8a and 8b represent the regression performed using the statistical methods of linear and polynomial to the order of 2. Figures 8c, 8d, 8e, and 8f represent the ML method for regression including SVR linear, polynomial, RBF and RF functions. Figures 8g and 8h represent the deep learning models that includes the DNN-MLP and LSTM functions. It shows that majority of the regression lines falls within the training dataset showing a positive correlation between the pipe length and the construction duration. The general fit of the regression models follows closely with the dataset used to produce the regression models. However, the confidence interval bands are far broader for the DL models.

The performance of each method was validated using statistic based evaluation metrics that includes: MSE, RMSE, MAE and R^2 as shown numerically in Table 8 and visually represented in Figure 9. The validation criteria show that, in general, ML models, RF presented best accuracy when compared to statistical methods and deep learning models. Out of the well performed ML models, RF regression function proved to be the most effective model for this specific application with a high R^2 value of 0.847, lowest MSE, MAE and RMSE of 0.024, 0.375 and 0.446, respectively. On the other hand, the worst performing model was SVR polynomial kernel with the lowest R^2 value of 0.749 and the highest MSE, MAE and RMSE of 0.706, 0.689 and 0.806, respectively. These results indicate a strong correlation between the selected models and the actual construction duration represented by the validation dataset.



a) Statistical linear regression

b) Statistical polynomial regression (n = 2)

c) SVR: linear kernel

d) SVR: polynomial kernel

e) SVR: radial basis function kernel

f) Random forest regressor

g) DNN-MLP regression model

h) LSTM regression model

Figure 8. Regression model results

The regression models were used to produce a predicted result against 17 validation dataset that had no influence in the regression models. The process was kept consistent by using pipe line length as the independent variables that served as an input for the regression models to make prediction on the construction duration. The extracted construction duration was compared to the real construction duration to produce prediction accuracy as presented in Table 9 where it was represented visually in Figure 10. In prediction, polynomial regression demonstrated the strongest performance with an average accuracy percentage of 98.79% whereas least accurate prediction was made with DNN-MLP model that scored 87.052% of average accuracy percentage. The rankings, based on the R^2 score, are as follows: RF, DNN-MLP, SVR RBF, polynomial regression, LSTM, Linear regression, SVR linear and SVR polynomial. However, according to the performing the prediction, the rankings are as follows: polynomial regression, SVR linear, linear regression, SVR RBF, RF, LSTM, SVR polynomial and DNN-MLP. In summary, the statistical criteria do not reflect the prediction capability of the regression models produced. This may be due to the statistical criteria deriving from the k-fold cross validation. While it removes the bias that may arise from a randomly split training and testing data, it may not necessarily reflect the prediction capability. In general, all of the regression models, bar SVR polynomial, has shown a positive correlation against the sewage construction data. However, from this study, it was found that a methodological variable that must be taken into consideration is the computation time. Statistic and ML methods generally performed significantly faster than the deep learning methods; as the methods did not require numerous training epochs and hyperparameter adjustment to find appropriate training and validation less and accuracy functions.

Table 8. Regression algorithm results score

Regression method	R^2 score	MSE	MAE	RMSE	Rank
Linear regression	0.823	0.264	0.392	0.474	6
Polynomial regression (n=2)	0.832	0.259	0.388	0.470	4
SVR linear	0.823	0.253	0.385	0.465	6
SVR polynomial	0.749	0.706	0.689	0.806	8
SVR radial basis function	0.834	0.244	0.384	0.463	3
Random forest regressor	0.847	0.224	0.374	0.446	1
DNN-MLP	0.841	0.254	0.394	0.475	2
LSTM	0.826	0.399	0.501	0.609	5

Table 9. Prediction against real data

Real length (Km)	Real construction time (Month)	Regression model construction time predictions (Months)							
		Linear Regression	Polynomial regression	SVR linear	SVR polynomial	SVR RBF	RF	DNN-MLP	LSTM
22	22	23.286	20.976	23.146	28.498	20.164	28.369	16.921	26.991
25	32	31.953	32.385	32.038	30.960	32.821	38.036	21.428	28.467
26	29	29.353	29.291	29.370	30.909	29.633	37.954	24.393	29.371
27	24	25.019	23.508	24.924	29.734	23.133	33.663	24.932	29.530
27	24	25.019	23.508	24.924	29.734	23.133	24.058	25.875	29.818
28	29	29.353	29.291	29.370	30.909	29.633	29.668	27.277	30.343
29	30	30.219	30.354	30.259	30.947	30.757	28.369	29.022	31.175
29	30	30.219	30.354	30.259	30.947	30.757	34.213	30.126	31.779
30	30	30.219	30.354	30.259	30.947	30.757	28.369	30.715	32.087
32	36	35.420	36.072	35.595	31.186	36.247	31.750	36.235	34.610
32	38	37.153	37.727	37.373	31.662	37.643	31.750	36.235	34.610
35	38	37.153	37.727	37.373	31.662	37.643	37.954	39.589	37.300
35	32	31.953	32.385	32.038	30.960	32.821	24.058	39.708	37.372
36	33	32.820	33.354	32.927	30.966	33.763	24.058	41.149	38.252
37	37	36.287	36.915	36.484	31.380	36.969	33.663	41.726	38.598
41	38	37.153	37.727	37.373	31.662	37.643	29.668	46.493	41.636
42	36	35.420	36.072	35.595	31.186	36.247	31.75	47.167	42.09
Average percentage accuracy (%)		98.139	98.790	98.406	87.965	97.693	93.658	87.052	89.303

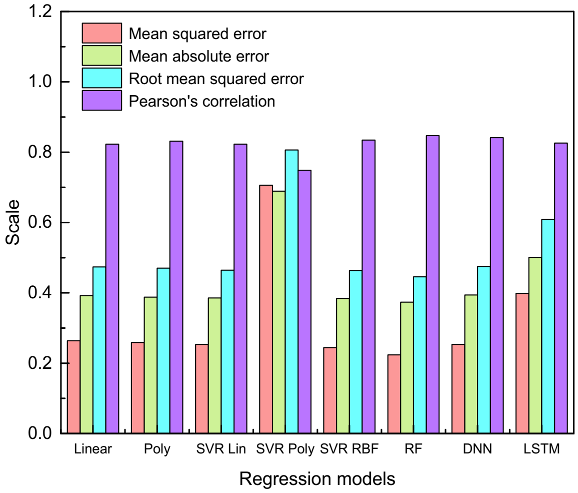


Figure 9. Regression model performance criteria results

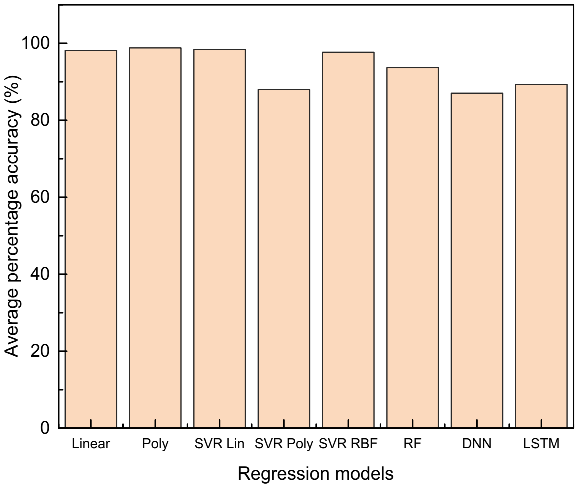


Figure 10. Average percentage accuracy for respective regression models against real data

5. Discussion

In this study, the methods for performing regression are based on previous literature where various studies have reported successful use of ML models in detecting and producing correlation between input and output with smaller data sets (Zhang et al., 2021). ML models have also shown capability to be able to identify data patterns and transform historical data to support a decision-making system (Awad & Khanna, 2015). In other studies, findings demonstrated that DL models such as DNN have better accuracy than ML models; due to their high ability to learn from training data (Darko et al., 2023; Wang et al., 2022). Although previous studies had mentioned the benefits of using DL models, the results from this study shows that this is not necessarily true. Although, ML and DL methods have shown excellence in their respective tasks from previous studies, traditional numerical methods cannot be ignored from perspective of prediction accuracy and computation time. This shows a potential in further research to develop and optimize ML and DL models customized for a specific case such as sewage pipeline construction. Moreover, contradictory findings emphasize the significance of data availability and contextual factors over model accuracy, complexity, and estimation processes (Amoore, 2023). The issue may even arise from hyperparameter optimization methods. In this study, a well-known and reliable optimization strategy was chosen to optimize the hyperparameters despite the development of other optimization strategies, such as grey wolf optimization and ant lion optimization (Mahmoodzadeh et al., 2022c; Nair et al., 2024), due to its commercial availability that signifies its validity.

DL method is a trainable model based on previous data that is based on a neural network with hidden layers and number of neurons as its main variables. In this study, a customizable model called DNN-MLP is used where optimized number of hidden layers as the input against a relatively simple set of data. The approach is holistically different to ML and statistical method. Theoretically, the DL method should fully learn the pattern between the dependent and independent variable that produces the most reliable result. However, this method is dependent on a large amount of training data, accurate hyperparameters and sufficient computing power to properly execute. On the other hand, LSTM was chosen as it is an openly and commercially available deep learning prediction method. It's a method where memory is retrained as it includes memory cells. Similar to the DNN-MLP method, this method requires a tremendous amount of training data and sufficient computing power to be executed.

The results of the regression showed a more than acceptable line of best fit across the numerous regression method conducted. Two sets of results were performed where statistical evaluation criteria were used to make an initial evaluation and a prediction was made against a validation dataset, that had no influence in the training and producing the regression models. The validation is conducted to make a prediction on construction duration depending on the sewage pipeline length. The predicted construction duration was then compared to the real construction duration in order to achieve an average of prediction accuracy. For the former, statistical evaluation criteria, RF regressor produced the best result where Pearson's correlation of 0.847 was calculated. The MAE, MSE and RMSE reflected the Pearson correlation where it was calculated to be 0.224, 0.374 and 0.446, respectively. However, despite the more reliable statistical evaluation criteria, it ranked fifth when predicting the construction duration against a real pipe length. This shows that for ML and deep learning methods, statistical evaluation criteria may not be the most suitable choice for a regression method with a small amount of dataset.

As mentioned previously, the results from this study shows that statistical performance criterions do not necessary reflect the prediction capability of the regression models whether that be derived from numerical, ML or DL methods despite the celebrated research trend in the latter. Moreover, computation time for the DL methods that included the hyperparameter optimization far exceeded than those of statistical and ML processes. This begs the question: is the state-of-the-art method necessarily better than its predecessor for specific cases such as: predicting sewage pipeline construction duration? Especially for those dealing with a simpler form of numerical data, could modern methods be overpowered for more simpler tasks with simpler inputs? Are state-of-the-art regression frameworks reliable in that it is capable of producing accurate and reliable results?

Upon observation of the graphs in Figure 8, majority of the validation data falls closely to the regression line produced by the respective methods. In general, for deep learning methods, observation showed that validation datasets fell within the regressed line of best fit for shorter sewage pipe length. This was deemed to be thought due to the amount of available data for training purposes. Therefore, it shows potential to be used when sufficient amount of data is available.

Analysis and validation show, RF, DNN-MLP and SVR RBF kernel and the random forest model demonstrated the highest predictive accuracy, while statistical linear regression yielded the lowest accuracy. As for the deep learning models, they demonstrated low accuracy in comparison with the machine learning models. This suggests that novel machine learning and deep learning methods may not necessarily reflect the best outcome when compared to statistical methods. It is essential that appropriate regression method is chosen by specialists with knowledge regarding: construction, management and numerical analysis in order to provide a more reliable regression outcome to those in management responsible for planning the construction determining the outcome of project success.

6. Conclusions

Given the scarcity of sewage pipeline data, numerous methods based on previous studies needs to be explored and compared to sustain the developing technologies in the field of machine learning and deep learning. Providing reliable data for project and construction managers within civil engineering sector is significant that take into account for: environmental risk, health risk and traffic disruption. In order to minimize the mentioned and unmentioned risks, predicting construction duration, that is dependent on the scale, of sewage pipeline construction is essential.

This research is a methodological based study that compared methods suggested by previous works and was evaluated using statistical evaluation method and validated by predicting actual construction. 83 previous successful sewage pipeline construction data is collected where dependent and independent variable was extracted according to correlation against pipe line length; that determines the scale of the construction. The 83 datasets were separated to training and validation dataset with the ratio of 8:2. The training dataset was strictly used to produce the regression models using statistical regression of linear and polynomial in the order to 2. The same dataset was used to perform 4 machine learning and 2 deep learning regression methods called: SVR linear, polynomial, RBF, RF, DNN-MLP and LSTM.

The regression models, upon observation, showed generally reliable results where majority of validation datasets falls within the boundaries of the confidence interval. From further analysis, statistical evaluation criteria of MSE, MAE and RMSE was performed with high scores. Moreover, prediction against the validation dataset was performed where sewage pipeline length was used a dependent variable to extract the construction duration that outperformed other ML and DL based methods.

The novelty of this study comes from the small amount of available data collected where only few studies have performed similar study with sufficient amount of data on a particular case such as the sewage pipeline construction. The results from this study highlights the need for more development in the ML and DL methods of performing numerical management tasks with a limited amount of available data. Though there are a wide existing body of knowledge on topic of construction duration prediction for various fields. While the state-of-the-art methods are celebrated in terms of accuracy and reliability, the findings from this study shows otherwise. The contradicting result may arise from various sources such as: data preprocessing, hyperparameter optimization, cross validation or even the regression functions. There is a need for further study that explores this to improve the ML and DL framework to achieve higher accuracy and reliability. Furthermore, in terms of computation time, ML and DL has been the least efficient in performing calculations. Therefore, state-of-the-art methods that are heavily dependent on the computation power may not be most suitable method for obtaining reliable results for this particular problem.

The results in this paper shows that among the regression techniques employed, the RF regressor achieved the highest R^2 score of 0.847 that averaged from the k-fold cross validation. On the other hand, SVR polynomial demonstrated the lowest R^2 score at 0.749 showcasing its comparatively low ability to forecast accurately. Moreover, when comparing the predicted results against the real data, the polynomial regression model demonstrated highest predictive capabilities achieving an average percentage accuracy of 98.79%. Moreover, both the SVR linear and linear regression models consistently provided estimations of construction time that closely matched the actual data, showing an average percentage accuracy of approximately 98.406% and 98.139%, respectively. The inaccuracy; however, of the DNN-MLP presented an average percentage accuracy of 87.052%, which reveals their limited precision in predicting construction durations compared to machine learning models. In summary, the results indicate that numerical methods of producing linear and polynomial regression cannot be ignored while most of the focus and attention in academic studies are geared towards ML and DL methods. While the ML and DL methods are capable in making reliable predictions, it may not necessarily be the correct choice when taking into account for computation time and power requirements also.

The contribution to this study is as follows: first, it showcased that ML and DL methods are capable of reliably predicting construction duration for a sufficient amount of dataset. Second, statistical processes have outperformed the novel ML and DL methods highlighting the outdated methods still needs to be taken into consideration depending on the problem and data available. Third, there is a need to explore and develop even more methods that improves the general ML and DL based frameworks by investigating alternative hyperparameter optimization algorithms and cross validation methods.

According to the results of this research, using statistic-based regression in future prediction of the construction duration of sewage pipelines can outperform modern methods deriving from ML and DL. Furthermore, more research exploring the prediction of the construction costs in the field of sewage pipeline using: different frameworks, hyperparameter optimization and cross validation methods is necessary to improve the ML and DL frameworks.

The limitation of this study comes in two folds of exploration and availability in data. While there are conventional and well-known optimization algorithms, based on evolutionary and numerical processes, there are still emerging methods that needs to be explored such as Antlion optimization and Grey Wolf optimization to name a few (Nair et al., 2024). As the information age progresses, studies in exploring as much available options as possible is necessary with the aim obtaining the best result possible as is the primary aim of optimization. Moreover, optimization parameter needs to be explored further as this study has opted for a conservative 500 generations in the optimization strategy. Although this study has provided sufficient data to allow for a fairly reliable result, there are still plenty of room for improvement. With more data in more diverse situations for a specific purpose, ML and DL techniques shows potential in forecasting project variables that influences project planning. Moreover, this study was only performed on available data and computation power that neglected other influential factors that may hinder the project outcome such as: weather and environment conditions, financial resources, and labor availability.

Abbreviations

ANN –	Artificial neural network;
BPNN –	Back-propagation neural network;
DT –	Decision tree;
DL –	Deep learning;
DNN –	Deep neural network;
XGBOOST –	Extreme gradient boosting;
GPR –	Gaussian process regression;
KNN –	K-nearest neighbor;
LR –	Logistic regression;
LSTM –	Long short-term memory;
ML –	Machine learning;
MAE –	Mean absolute error
MSE –	Mean squared error;
MLP –	Multiple layer perceptron;
MLR –	Multiple linear regression;
NSGA-II –	Non-dominated sorting genetic algorithm;
N/A –	Not available;
R^2 –	Pearson's correlation;
RBF –	Radial basis function;
RF –	Random forest;
RNN –	Recurrent neural network;
RMSE –	Root mean squared error;
SVR –	Support vector regression.

Acknowledgements

The authors would like to thank the editors and the reviewers for their constructive and valuable comments and careful reading, which significantly improved the paper.

Funding

This research received no external funding.

Author contributions

People who contributed to the work are listed in this section along with their contributions: Supervision, JHK; data collection, KYL, NN; statistical analysis, SJP; writing original draft of the article, NN and SJP; review and editing of manuscript, JHK and SJP.

Disclosure statement

Authors do not have any competing financial, professional, or personal interests related to other parties.

References

Abed, Y. G., Hasan, T. M., & Zehawi, R. N. (2022). Cost prediction for roads construction using machine learning models. *International Journal of Electrical and Computer Engineering Systems*, *13*(10), 927–936. https://doi.org/10.32985/ijeces.13.10.8

Abu Hammad, A. A., Ali, S. M. A., Sweis, G. J., & Sweis, R. J. (2010). Statistical analysis on the cost and duration of public building projects. *Journal of Management in Engineering*, *26*(2), 105–112. https://doi.org/10.1061/(ASCE)0742-597X(2010)26:2(105)

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., & Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, *32*, Article 101827. https://doi.org/10.1016/j.jobe.2020.101827

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme gimpoiradient boosting-based machine learning approach for green building cost prediction. *Sustainability*, *14*(11), Article 6651. https://doi.org/10.3390/su14116651

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), Article 53. https://doi.org/10.1186/s40537-021-00444-8

Amoore, L. (2023). Machine learning political orders. *Review of International Studies*, *49*(1), 20–36. https://doi.org/10.1017/S0260210522000031

Awad, M., & Khanna, R. (2015). *Efficient learning machines. Theories, concepts, and applications for engineers and system designers*. Springer. https://doi.org/10.1007/978-1-4302-5990-9

Baloyi, L., & Bekker, M. C. (2011). Causes of construction cost and time overruns: The 2010 FIFA World Cup stadia in South Africa. *Acta Structilia*, *18*(1), 51–67.

Barjouei, H. S., Ghorbani, H., Mohamadian, N., Wood, D. A., Davoodi, S., Moghadasi, J., & Saberi, H. (2021). Prediction performance advantages of deep machine learning algorithms for two-phase flow rates through wellhead chokes. *Journal of Petroleum Exploration and Production Technology*, *11*(3), 1233–1261. https://doi.org/10.1007/s13202-021-01087-4

Bayram, S., Ocal, M. E., Laptali Oral, E., & Atis, C. D. (2016). Comparison of multi layer perceptron (MLP) and radial basis function (RBF) for construction cost estimation: the case of Turkey. *Journal of Civil Engineering and Management*, *22*(4), 480–490. https://doi.org/10.3846/13923730.2014.897988

Behnia, D., Ahangari, K., Noorzad, A., & Moeinossadat, S. R. (2013). Predicting crest settlement in concrete face rockfill dams using adaptive neuro-fuzzy inference system and gene expression programming intelligent methods. *Journal of Zhejiang University SCIENCE A*, *14*(8), 589–602. https://doi.org/10.1631/jzus.A1200301

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of The Total Environment*, *721*, Article 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dang-Trinh, N., Duc-Thang, P., Cuong, T. N.-N., & Duc-Hoc, T. (2023). Machine learning models for estimating preliminary factory construction cost: case study in Southern Vietnam. *International Journal of Construction Management*, *23*(16), 2879–2887. https://doi.org/10.1080/15623599.2022.2106043

Darko, A., Glushakova, I., Boateng, E. B., & Chan, A. P. C. (2023). Using machine learning to improve cost and duration prediction accuracy in green building projects. *Journal of Construction Engineering and Management*, *149*(8), Article 04023061. https://doi.org/10.1061/JCEMD4.COENG-13101

DataFlair. (2022). *Advantages and disadvantages of machine learning language*. https://data-flair.training/blogs/advantages-and-disadvantages-of-machine-learning/

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Dong, J., Chen, Y., & Guan, G. (2020). Cost index predictions for construction engineering based on LSTM neural networks. *Advances in Civil Engineering*, *2020*, Article 518147. https://doi.org/10.1155/2020/6518147

Doyle, M. W., & Havlick, D. G. (2009). Infrastructure and the environment. *Annual Review of Environment and Resources*, *34*(1), 349–373. https://doi.org/10.1146/annurev.environ.022108.180216

Fang, C., Zhang, X., Cheng, Y., Wang, S., Zhang, L., & Yang, Y. (2019). Fault diagnosis for brake system in high-speed trains using the phased features and multi-layer perceptron. *IOP Conference Series*: *Materials Science and Engineering*, *470*, Article 012007. https://doi.org/10.1088/1757-899X/470/1/012007

Ganiyu, B., & Zubairu, I. (2010). Project cost prediction model using principal component regression for public building projects in Nigeria. *Journal of Building Performance*, *1*(1), 21–28.

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2012). An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. *GIScience* & *Remote Sensing*, *49*(5), 623–643. https://doi.org/10.2747/1548-1603.49.5.623

Gujar, R., & Vakharia, V. (2019). Prediction and validation of alternative fillers used in micro surfacing mix-design using machine learning techniques. *Construction and Building Materials*, *207*, 519–527. https://doi.org/10.1016/j.conbuildmat.2019.02.136

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Idowu, O. S., & Lam, K. C. (2020). Conceptual quantities estimation using bootstrapped support vector regression models. *Journal of Construction Engineering and Management*, *146*(4), Article 04020018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001780

Kamali, M. Z., Davoodi, S., Ghorbani, H., Wood, D. A., Mohamadian, N., Lajmorak, S., Rukavishnikov, V. S., Taherizade, F., & Band, S. S. (2022). Permeability prediction of heterogeneous carbonate gas condensate reservoirs applying group method of data handling. *Marine and Petroleum Geology*, *139*, Article 105597. https://doi.org/10.1016/j.marpetgeo.2022.105597

Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merchán, E. C., Branke, J., & Bischl, B. (2023). Multi-objective hyperparameter optimization in machine learning – An overview. *ACM Transactions on Evolutionary Learning and Optimization*, *3*(4), Article 16. https://doi.org/10.1145/3610536

Khedr, A. M., Arif, I., P V Raj, P., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting*, *Finance and Management*, *28*(1), 3–34. https://doi.org/10.1002/isaf.1488

Kim, K.-Y. (2022). *Old water and sewage pipes – land collapse caused by poor construction*. Donga Ilbo (in Korean). http://www.donga.com/news/Opinion/article/all/20220112/111202406/1

Kim, Y.-J., Yeom, D.-J., & Kim, Y. S. (2019). Development of construction duration prediction model for project planning phase of mixed-use buildings. *Journal of Asian Architecture and Building Engineering*, *18*(6), 586–598. https://doi.org/10.1080/13467581.2019.1696207

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations* (*ICLR 2015*), San Diego, CA, USA.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lee, H., Chung, S.-H., & Choi, E.-J. (2016). A case study on machine learning applications and performance improvement in learning algorithm. *Journal of Digital Convergence*, *14*(2), 245–258. https://doi.org/10.14400/JDC.2016.14.2.245

Lin, M.-C., Tserng, H. P., Ho, S.-P., & Young, D.-L. (2011). Developing a construction-duration model based on a historical dataset for building project. *Journal of Civil Engineering and Management*, *17*(4), 529–539. https://doi.org/10.3846/13923730.2011.625641

Maclin, R., & Opitz, D. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Mahmoodzadeh, A., & Zare, S. (2016). Probabilistic prediction of expected ground condition and construction time and costs in road tunnels. *Journal of Rock Mechanics and Geotechnical Engineering*, *8*(5), 734–745. https://doi.org/10.1016/j.jrmge.2016.07.001

Mahmoodzadeh, A., Mohammadi, M., Daraei, A., Rashid, T. A., Sherwani, A. F. H., Faraj, R. H., & Darwesh, A. M. (2019). Updating ground conditions and time-cost scatter-gram in tunnels during excavation. *Automation in Construction*, *105*, Article 102822. https://doi.org/10.1016/j.autcon.2019.04.017

Mahmoodzadeh, A., Mohammadi, M., Daraei, A., Farid Hama Ali, H., Ismail Abdullah, A., & Kameran Al-Salihi, N. (2021). Forecasting tunnel geology, construction time and costs using machine learning methods. *Neural Computing and Applications*, *33*(1), 321–348. https://doi.org/10.1007/s00521-020-05006-2

Mahmoodzadeh, A., Mohammadi, M., Abdulhamid, S. N., Ibrahim, H. H., Ali, H. F. H., Nejati, H. R., & Rashidi, S. (2022a). Prediction of duration and construction cost of road tunnels using Gaussian process regression. *Geomechanics and Engineering*, *28*(1), 65–75. https://doi.org/10.12989/gae.2022.28.1.065

Mahmoodzadeh, A., Nejati, H. R., & Mohammadi, M. (2022b). Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects. *Automation in Construction*, *139*, Article 104305. https://doi.org/10.1016/j.autcon.2022.104305

Mahmoodzadeh, A., Nejati, H. R., Mohammadi, M., Hashim Ibrahim, H., Rashidi, S., & Ahmed Rashid, T. (2022c). Forecasting tunnel boring machine penetration rate using LSTM deep neural network optimized by grey wolf optimization algorithm. *Expert Systems with Applications*, *209*, Article 118303. https://doi.org/10.1016/j.eswa.2022.118303

Mahmoodzadeh, A., Taghizadeh, M., Mohammed, A., Ibrahim, H., Samadi, H., Mohammadi, M., & Rashidi, S. (2022d). Tunnel wall convergence prediction using optimized LSTM deep neural network. *Geomechanics and Engineering*, *31*(6), 545–556. https://doi.org/10.12989/gae.2022.31.6.545

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, *13*(3), Article e0194889. https://doi.org/10.1371/journal.pone.0194889

Malek Mohammadi, M., Najafi, M., Kaushal, V., Serajiantehrani, R., Salehabadi, N., & Ashoori, T. (2019). Sewer pipes condition prediction models: A State-of-the-art review. *Infrastructures*, *4*(4), Article 64. https://doi.org/10.3390/infrastructures4040064

Marino, D. L., Amarasinghe, K., & Manic, M. (2016). Building energy load forecasting using deep neural networks. In *42nd Annual Conference of the IEEE Industrial Electronics Society* (*IECON 2016*) (pp. 7046–7051). IEEE. https://doi.org/10.1109/IECON.2016.7793413

MathWorks. (2023). *What is machine learning?*. https://ww2.mathworks.cn/en/discovery/machine-learning.html?s_tid=srchtitle_Machine%20learning_3

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*, *4*(2), 103–112. https://doi.org/10.22115/SCCE.2020.226883.1205

Moret, Y., & Einstein, H. H. (2016). Construction cost and duration uncertainty model: Application to high-speed rail line project. *Journal of Construction Engineering and Management*, *142*(10), Article 05016010. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001161

Munns, A., & Bjeirmi, B. (1996). The role of project management in achieving project success. *International Journal of Project Management*, *14*(2), 81–87. https://doi.org/10.1016/0263-7863(95)00057-7

Nair, P., Vakharia, V., Shah, M., Kumar, Y., Woźniak, M., Shafi, J., & Fazal Ijaz, M. (2024). AI-driven digital twin model for reliable lithium-ion battery discharge capacity predictions. *International Journal of Intelligent Systems*, *2024*, Article 185044. https://doi.org/10.1155/2024/8185044

Obradović, D. (2017). The impact of tree root systems on wastewater pipes. In *Zajednički temelji 2017 – Peti skup mladih istraživača iz područja građevinarstva i srodnih tehničkih znanosti – Zbornik radova* (pp. 65–71). https://doi.org/10.5592/CO/ZT.2017.03

Obradović, D., Šperac, M., & Marenjak, S. (2023). Challenges in sewer system maintenance. *Encyclopedia*, *3*(1), 122–142. https://doi.org/10.3390/encyclopedia3010010

Opila, M. C. (2011). *Structural condition scoring of buried sewer pipes for risk based decision making* [PhD thesis]. University of Delaware, Newark, Delaware.

Peiman, F., Khalilzadeh, M., & Shahsavari-Pour N., & Ravanshadnia, M. (2025). Estimation of building project completion duration using a natural gradient boosting ensemble model and legal and institutional variables. *Engineering*, *Construction and Architectural Management, 32*(4), 2069–2104. https://doi.org/10.1108/ECAM-12-2022-1170

Pesko, I., Mucenski, V., Seslija, M., Radovic, N., Vujkov, A., Bibic, D., & Krkljes, M. (2017). Estimation of costs and durations of construction of urban roads using ANN and SVM. *Complexity*, *2017*, Article 450370. https://doi.org/10.1155/2017/2450370

Pierdzioch, C., & Risse, M. (2020). Forecasting precious metal returns with multivariate random forests. *Empirical Economics*, *58*(3), 1167–1184. https://doi.org/10.1007/s00181-018-1558-9

Popescu, M.-C., Balas, V., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*(7), 579–588.

Rafiei, M. H., & Adeli, H. (2018). Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, *144*(12), Article 04018106. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001570

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis*: *A research tool* (Springer texts in statistics). Springer-Verlag New York Inc.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning internal representations by error propagation. *Readings in Cognitive Science*. *A Perspective from Psychology and Artificial Intelligence*, 399–421. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2

Ministry of Environment Domestic Sewage Division. (2021). *2021 Sewerage statistics* (in Korean).

Saeidlou, S., & Ghadiminia, N. (2024). A construction cost estimation framework using DNN and validation unit. *Building Research* & *Information*, *52*(1–2), 38–48. https://doi.org/10.1080/09613218.2023.2196388

Shoar, S., Chileshe, N., & Edwards, J. D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, *50*, Article 104102. https://doi.org/10.1016/j.jobe.2022.104102

Son, H., & Kim, C. (2015). Early prediction of the performance of green building projects using pre-project planning variables: Data mining approaches. *Journal of Cleaner Production*, *109*, 144–151. https://doi.org/10.1016/j.jclepro.2014.08.071

Sueri, M., & Erdal, M. (2022). Early estimation of sewerage line costs with regression analysis. *Gazi University Journal of Science*, *35*(3), 822–832. https://doi.org/10.35378/gujs.949726

Taye, M. M. (2023). Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers*, *12*(5), Article 91. https://doi.org/10.3390/computers12050091

Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. *SN Applied Sciences*, *2*(10), Article 1703. https://doi.org/10.1007/s42452-020-03497-1

Vakharia, V., & Gujar, R. (2019). Prediction of compressive strength and Portland cement composition using cross-validation and feature ranking techniques. *Construction and Building Materials*, *225*, 292–301. https://doi.org/10.1016/j.conbuildmat.2019.07.224

Wang, R., Asghari, V., Cheung, C. M., Hsu, S. C., & Lee, C. J. (2022). Assessing effects of economic factors on construction cost estimation using deep neural networks. *Automation in Construction*, *134*, Article 104080. https://doi.org/10.1016/j.autcon.2021.104080

Yan, X., & Su, X. G. (2009). *Linear regression analysis*. *Theory and computing*. World Scientific. https://doi.org/10.1142/6986

Yeom, D.-J., Seo, H.-M., Kim, Y.-J., Cho, C.-S., & Kim, Y. (2018). Development of an approximate construction duration prediction model during the project planning phase for general office buildings. *Journal of Civil Engineering and Management*, *24*(3), 238–253. https://doi.org/10.3846/jcem.2018.1646

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. *Neural Computing and Applications*, *32*, 1609–1628. https://doi.org/10.1007/s00521-019-04212-x

Yu, P.-S., Chen, S.-T., & Chang, I.-F. (2006). Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, *328*(3–4), 704–716. https://doi.org/10.1016/j.jhydrol.2006.01.021

Yuan, J., Chen, W., Tan, X., Yang, D., & Wang, S. (2019). Countermeasures of water and mud inrush disaster in completely weathered granite tunnels: A case study. *Environmental Earth Sciences*, *78*(18), Article 576. https://doi.org/10.1007/s12665-019-8590-8

Zakaria, Z., Ismail, S., & Yusof, A. (2012). Cause and impact of dispute and delay the closing of final account in Malaysia construction industry. *Journal of Southeast Asian Research*, *2012*, Article 975385. https://doi.org/10.5171/2012.975385

Zhang, S., & Li, X. (2024). A comparative study of machine learning regression models for predicting construction duration. *Journal of Asian Architecture and Building Engineering, 23*(6), 1980–1996. https://doi.org/10.1080/13467581.2023.2278887

Zhang, X., Wang, Z., Liu, D., Lin, Q., & Ling, Q. (2021). Deep adversarial data augmentation for extremely low data regimes. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(1), 15–28. https://doi.org/10.1109/TCSVT.2020.2967419

Zheng, Z., Zhou, L., Wu, H., & Zhou, L. (2023). Construction cost prediction system based on Random Forest optimized by the Bird Swarm Algorithm. *Mathematical Biosciences and Engineering*, *20*(8), 15044–15074. https://doi.org/10.3934/mbe.2023674