Int. J. Nonlinear Anal. Appl. 13 (2022) 2, 2205–2218
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.27673.3684

Machine learning algorithms for constructions cost prediction: A systematic review

Yasamin Ghadbhan Abed[a,*], Taha Mohammed Hasan[a], Raquim Nihad Zehawi[b]

[a]Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq

[b]Department of Highway and Airport Engineering, College of Engineering, University of Diyala, Diyala, Iraq

(Communicated by Javad Vahidi)

Abstract

Machine learning plays a vital role in construction estimation which could make improve the project's safety, and reliability. Many studies have been proposed to explore the potential opportunities to review this technology in the construction cost in structure and transport fields. However, no comprehensive study to review the global research trends on this area's advancement in construction cost. The goal is to taxonomy, review, and summarize the state-of-the-art knowledge body on this topic in a systematic manner based on machine learning (ML) and deep learning (DL) approaches. To achieve this, this paper considered many studies in construction management related to bibliographic records retrieved from the Scopus database by adopting a quantitative analysis approach. This paper found that from 2017 to 2021, there has been a considerable increase in the number of publications in this domain. We categorized and explained civil projects into structures and transport cost, ML/DL as supervised and unsupervised approaches, and the evaluation metrics proposed to evaluate the performance of ML-Cost estimations in the civil area. The findings will help both professionals and researchers to understand and evolve the recent trend research ML/DL methodologies and their role played in the construction management domain.

Keywords: Systematic review, Construction, Cost estimation, Machine learning, Deep learning
2020 MSC: 68T07

1 Introduction

The prediction of conceptual costs represents one of the fundamental criteria in the projects' decision-making during the initial stages in the domains of civil engineering. The financial cost represents one of the key criteria to design, construct and maintain roads, bridges, buildings, canals, dams, airports, railroads, and so on. The prediction was used for the costs of the different construction domains; the purpose of the cost estimation is to predict the cost of the assumed values. Consequently, it is important to specify factors that have an essential effect on a project's cost for predicting the required cost of projects. The prediction should be concluded during a specified time. Thus, the precise estimation of conceptual costs represents a challenge that should be handled by decision-makers, project managers, and cost engineers [20]. Different approaches are proposed to calculate the cost of Roads and Bridges and highway projects. Early, the engineers or employers count manually the predicted or true estimation of project cost.

*Corresponding author
Email addresses: scicompms2133@uodiyala.edu.iq (Yasamin Ghadbhan Abed), dr.tahamh@uodiyala.edu.iq (Taha Mohammed Hasan), raquim_zehawi@uodiyala.edu.iq (Raquim Nihad Zehawi)

Received: April 2022 *Accepted:* June 2022

With the increase in the human race, the companies calculated the estimation of cost by limited software or tools like excel, access, SPSS, or static program. These programs are not capable of solving complex problems as prediction costs. Therefore, the effectiveness of the static approach contains high false positives or negatives, and it takes a long time to accomplish the task. Another approach is the dynamic technique of self-learning for solving problems, and it enables finding an optimal solution for complex problems [17]. Various Artificial intelligence (AI) models were utilized for predicting the estimated cost of construction domain research. In some of these prediction models, a kind of simple linear regression model named Ordinary Least Square (OLS) is utilized for forecasting the future cost of construction projects, and the obtained results demonstrate that these models provide prediction accuracies between 91% and 97%. There are several models of Artificial Neural Networks (ANNs) were recently proposed for predicting the estimated cost of civil engineering research, some researchers found that the ANNs model has the capability of predicting the structure works cost of highway projects with 93.19% of accuracy, while other researchers utilized other ANNs models such as RBFNN, GRNN, MLP, etc. to estimate the cost of road construction and the obtained accuracy was 95%. In bridge construction projects, some researchers utilized the support vector machines (SVM) method to estimate the cost, and the accuracy was 98%. Although several algorithms are used to build ML/Deep learning (DL) models for the estimated cost, few systematic reviews or survey studies are focused on reviewing the ML/DL models with estimation cost in construction domains research as shown in Table 1.

Table 1: Some systematic reviews or survey studies concerning ML/DL models with estimation cost in construction domains.

Ref.	**Systematic Review**	**ML approach**	**Accuracy**	**Error**	**Fields**	**Years**	**covered studies**
[34]	✗	Ms.	✓	✓	Highway	2020	Evaluation ML algorithms
[23]	✓	ANN	✗	✗	Building	2017	Review Challenge of Nonparametric Models for construction projects
[51]	✓	MLs	✗	✗	Building	2021	Review Only
[5]	✓	ANN	✗	✗	Transport	2017	Review Only
[13]	✓	ML	✓	✓	Tunnels	2020	Comparative ML Algorithms For Field canals improvement projects (FCIPs)
[46]	✓	ML	✗	✗	Multi fields	2020	Systematic Review for the limited machine learning techniques for cost construction
Proposed work	✓	ML and DL	✓	✓	Multi Domains	2021	Comprehensive Systematic Review of ML and DL for cost construction

However, the presented review papers or surveys did not encompass all the research accomplished in the construction projects field of estimating cost and only the ones that included specified keywords in construction projects analysis with ML methods. In this study, no restriction is imposed on the kind of proposed works carried out on the subject and with up to date of publications as well. This research is a novel contribution to the survey of construction domains with an analysis of the effectiveness of ML and DL techniques based on criteria such as mean square error (MSE), accuracy, precision, etc., therefore, it is necessary to need systematic review for analysis performance of ML/DL for previous works of civil research based on certain factors, which are; the effectiveness rate, and the period project with the estimated final budget. The rest of this review is formed as follows; Section 2 abbreviated some recently existing reviews concerning AI in the field of civil engineering and basic information for ML or DL techniques; In section 3, the considered research methodology is described. In section 4, the research method presented in this survey is detailed. The article classification considering the research questions is offered in section 5. In section 6, a discussion of the performance analysis is provided with a concentration on the current gaps in this research. Finally, the main conclusions of the systematic review are presented in the last section 7.

2 Research Methodology

Based on the literature surveys, there is no comprehensive performance analysis using various models of AI concerning conceptual estimation cost. The fundamental aim of this survey is to evaluate the prediction accuracies of various AI models to show the best prediction models that obtained the highest results accuracies. Furthermore, this survey presents an extensive comparison of the AI models' performance for guiding practitioners and researchers in conceptual cost modeling. The most common AI models are considered in this survey such as random forest (RF), extreme gradient boosting machines (XGBoost), SVM, ANNs, Decision Tree (DT), Deep Neural Network (DNN), Conventional Neural Network (CNN), and Recurrent Neurons Network (RNN). This systematic review includes several stages as demonstrated in Figure 1.

Stage 1: which is to define the words query to determine the significant articles, which is concerned with machine learning techniques for cost estimation in the civil domain. In order to acquire pertinent papers, the databases are searched utilizing several keywords; ("Machine learning" AND "construction") or ("Machine learning" AND "Roads") or "Highways" or "Buildings" or "Bridges" or "Airports" or "Dams" or "Tunnels" or "Railroads" or "Towers" and (at least one of the words) "prediction" or "estimation" or "forecasting". These utilized keywords represent the most significant guidance in this field that can assist to attain pertinent papers. Therefore, there are no imposed restrictions on the publication state of the extracted research. Moreover, the source research engine is selected based on trusted databases such as IEEE, Springer, and Elsevier.

Stage 2: it is to download appropriate articles that are proposed ML models for civil domain from Google scholar from trust source engines between 2017-2021.

Stage 3: it is to read and summarize the selected articles to form quick information, which contains summary details for the project type, location or dataset, effectiveness, and ML techniques. This information is utilized for creating new knowledge to assess the analysis of the performance of current ML models. Besides, the current challenges and recommend the trend future are concluded for the AI methods for cost construction estimation.

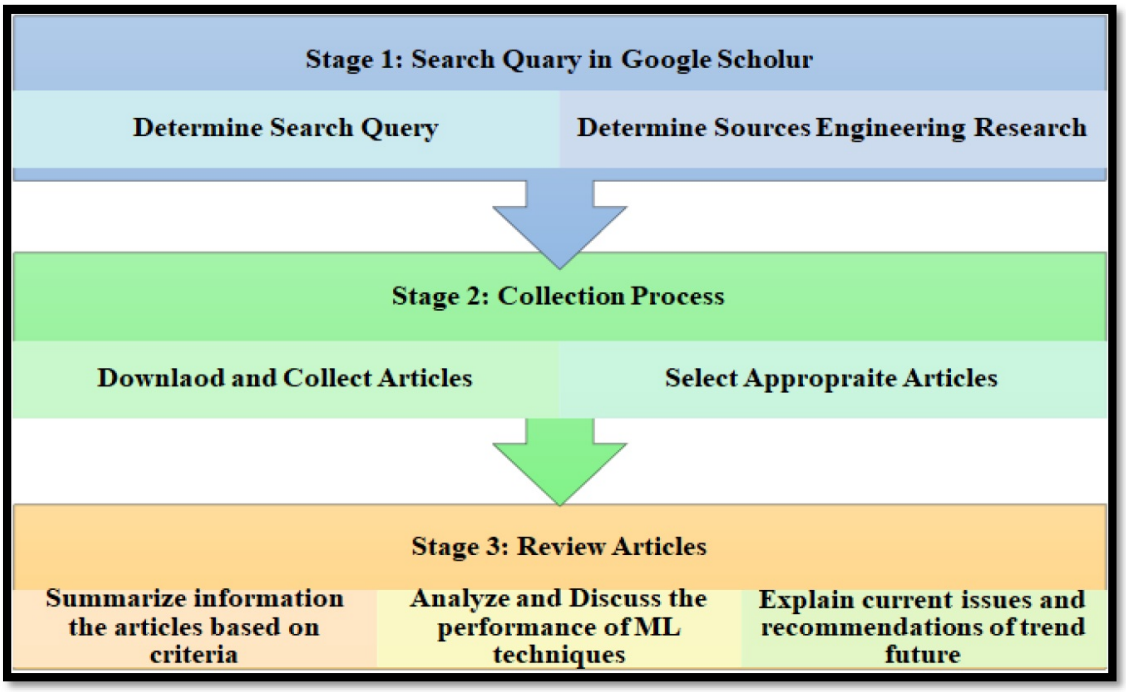

Figure 1: Methodology for systematic process.

2.1 Project Cost

Construction and management engineering is about creating things like bridges, buildings, roads, and railways. The purpose of engineering construction is to build these structures in the most efficient, safe, sustainable, and environmentally with low actual cost. This cost needs to be counted in accurate forms so that the owner of the project can make sure that there is value in return for money spent on the projected cost which can be divided based on the

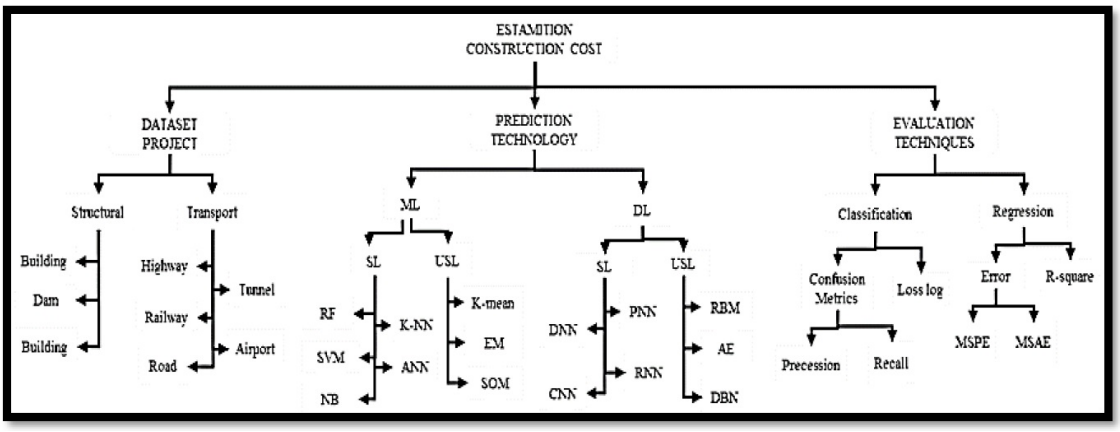

Figure 2: The taxonomy of cost estimation

type of project construction namely; Structural and transport cost construction. Structural cost is a specialty of civil engineering that focuses on building structures to endure the strains and forces of their environment while remaining safe, stable, and secure during its use, which includes:

- **Building Cost:** It is a structure having a roof and walls that remain in one location for the most part. Buildings are available in a wide range of sizes, shapes, and purposes. Buildings are classified into numerous categories, each of which has its own set of costs. In general, assumptions regarding the project's nature, as well as recognizing factors in all types of buildings, are critical. It is essential to calculate the infrastructure costs on a basis of per unit, utilizing comparable projects as a guide, with assumptions for different sorts of locations and unusual things. The costs of the unit might encompass a number of units of dwelling, rentable and parking areas, hospital beds, hotel bedrooms, prison cells, etc. The main variables for buildings projects prediction costs are building area, structure-envelope type, floor height, and the number of floors [37].
- **Bridge Cost:** This structure spans a physical obstacle without no inhibiting the path beneath it. It builds for allowing transit over the barrier that is usually somewhat that would be tricky or not possible to cross another way. Bridges can be categorized in a variety of ways, including the structural elements utilized, the loads they carry, whether they are fixed or moveable, and the materials they are made of. Bridges are classed as Beam, Truss, Cantilever, Arch, Suspension, and Cable-stayed bridges based on their structure. The main variables that are utilized for predicting the bridge construction cost are the kind of project, kind of bridge, material and structural solutions, supports' kinds and their foundation, class of load, the measures of basic size, the total width and length of decks, and the number of spans [15].
- **Tower cost:** Costs might influence the tower design chosen and, as a result, the project's final location. Understanding the financial implications of whether utilities are considering a self-supporting or guyed tower can bring much-needed insight. The most influential components in cost assessment are steel costs, foundation costs, erection expenses, land usage costs, and maintenance prices [39].
- **Dams Cost:** A dam is a structure that prevents or regulates the flow of surface or subsurface water. Dams are classed based on the dam type and size of the dam wall or reservoir, as well as the main purpose of the reservoir, which could be irrigation, hydropower generation, or water supply estimation. The entire cost of constructing a dam is determined by numerous elements; some authors base cost estimation and computation on the type of dam; the vast range in water supply reservoir results is likely to influence cost-effectiveness. According to the other authors, the cost is determined by the dam's position and closeness to the river. The estimation of the other author's cost depends on the location of the dam and its proximity to the riverbed also affects the reduction of the dam construction cost [39].

Transport cost deals with the type of civil engineering that deals with the planning, design, operation, and maintenance of transportation networks in order to help communities become smarter, safer, and more livable, which includes:

- **Road cost:** The road represents a broad way of transporting from one location to another, typically a ready surface that vehicles are able to utilize. The four major road functional classifications are interstates, collectors,

local roads, as well as other arterials. In the road construction costs, there were three main groups of factors, specifically, environmental factors, project factors, and the factors related to technical conditions of the project. These groups analyze the kind and scope of the road project, construction duration, width and length of the road contract, market situations of local construction, costs of actual construction, the energy situations, and macroeconomic indicators [8].

- **Highway cost:** A highway is a multi-lane route with a lot of traffic. Highways were created to connect cities and villages, and because they're wide and have high-speed limits, they cut travel time in half. The cost of construction for highways depends on several variables, including but not limited to the specific features of the project, macroeconomic indicators, local and regional construction market indicators, and energy market indicators [4].
- **Railroad cost:** a persistent road for automobiles track or equipment propelled via motors of self-contained or drawn via locomotives, including a rails' line fastened with ties on a roadbed. It's difficult, if not impossible, to predict the exact expenses of railway construction. However, being able to make realistic and representative estimates based on influencing elements that will affect the railway's cost will allow planning professionals to execute a larger range of analyses and, as a result, meet decision-making deadlines for network and service demands in less time. Cost estimation in railroads is based on the railway's design and construction (Design and Build) and raw materials [50].
- **Airports cost:** It is a location with buildings and passenger facilities where planes land and take off. On the basis of take-off and landing, geometric design, aircraft approach speed (FAA), and function, airports can be divided into four categories. Airport Construction refers to the construction, rebuilding, enlargement, relocation, maintenance, and repair of the many structures, infrastructure, and facilities on the Airport [48].
- **Tunnel cost:** It is an underground passage mined through the around rock/earth/soil and enclosed excepting in entry and departure [14]. There are many factors that affected the construction cost for the tunnel some authors used, the rock mass rating is a geology parameter that should be considered that can describe the tunnel's ground conditions and affect tunnel construction time and costs [51].

3 Prediction Technology

The algorithms of ML or DL encompass unsupervised learning and supervised learning. The algorithms of unsupervised learning detect unlabeled instances of data using the clustering learning method, and encompass; Expectation-Maximization (EM) clustering, K-means clustering, and Self Organizing Map (SOM). While the algorithms of supervised learning implement the detection (classification) depending on the labeled instances of data in the stage of training, and encompass SVM, DT, RF, Naive Bayes (NB), K-Nearest Neighbors (K-NN), and ANN, as demonstrated in Figure 2. Most algorithms used for cost estimation are described in the next sub-sections.

3.1 ML Approaches

A comprehensive categorization of AI models is presented in this sub-section. These models are depicted in detail and their applicability as cost predictors. There are lots of ML classifiers; which are depicted as follows [32]:

- **ANN:** In visual terms, it can represent as a weighted directed graph that includes several nodes (artificial neurons) and edges (directed edges with weights). The neurons take the input as a vector that is analogous to a specific pattern or image. The neuron's output utilizes as the input for other neurons, and the weights are modified during the ANN training for addressing the issues of classification. Typically, the architecture of ANN includes an input layer, an output layer, and multiple hidden layers that are completely or partially linked, each of which comprises neurons. The hidden layers work as an intermediary between the input layer and output layer and adjust the input in a specific manner for utilizing the output layer.
- **DT:** It frequently utilizes the algorithms of supervised learning for solving the issues of ML classifiers. The models of trees (classification trees) are efficiently utilized in the conditions that the target variable can take discrete values as inputs. The DT components include nodes, branches, and leaves. The leaves indicate the class labels and the branches indicate the attributes' set that produces the class labels. DT works on dividing the samples into several homogeneous collections in accordance with an ultimate significant divider in input determinatives. But, this model faces an issue of overfitting that is coped with by performing bagging and boosting methods.

- **SVM:** It specifies a hyper-plane that classifies the instances of training into multi or binary classification. Regarding, the SVM model takes the mentioned instances and related outputs, involving binary or N-array. Then, this model is built for categorizing the new instances. The instances of training are mapped into various points in the space of coordinate, these instances are linearly separated as input sets. Many hyper-planes that are capable of separating the instance sets of training are existing for selection. But, the appropriate selection represents the maximum distance from the nearest instance regarding any category. The model of SVM can be effectively implemented in the spaces of high-dimensional as well.

- **RF:** It effectively resolves the issue of overfitting that faces the DT model via utilizing the average of multiple DTs. This model represents an algorithm of ensemble learning that rectifies the regression and classification problems. RF requires multiple DTs construction during the timespan of training. It produces the mode of classes for a certain DT when performing a function of classification, and outputs more outstanding results compared with DT.

- **K-NN:** In accordance with [14], K-NN represents an algorithm of instances-based learning and a model of classification whose basics concentrate on its function of distance that calculates the differences or correlations among pairs of points or instances. Furthermore, in the K-NN model, there are several alternative measures of distance that can be used, comprising Euclidean distance as follows:

$$\left(D(a,b) = \sqrt{\sum_{i=1}^{r} (ai - bi)^2} \right) \tag{3.1}$$

  Where r indicates the whole dataset features quantity; ai indicates the i^{th} featured element of 'a' instance, and bi indicates the i^{th} featured element of 'b' instance. The nonparametric measure named Euclidean distance doesn't have any basic data dissemination conjecture. The model construction is determined from the dataset, and this model has proven to be useful since a significant portion of the data is derived from realistic datasets with no abiding with mathematical guesswork. While a lazy algorithm requires a model construction that doesn't entail the data points training. The entire data of training are used within the stage of testing, which is slower than the stage of training. In the worst-case scenarios, K-NN needs extra time, storage of training data for scanning the whole points of data.

- **NB:** this model is utilized as a probabilistic classifiers group that is generated via performing the theorem of Bayes. These models take the independence naive conjectures between every pair of attributes and features. In the stage of the training data processing, the NB is capable of competing with the most developed models within its field, like ANN and SVM. It is easily trained via utilizing a structure of supervised learning. In various actual implementations, the technique of maximal probability is implemented for calculating the NB models' parameters. Specifically, this model is capable of functioning with the Bayesian probability refusal or via utilizing any technique of Bayesian. The theorem of Bayes is given as follows:

$$\left(\rho(A|B) = \frac{\rho(A|B)\rho(A)}{\rho(B)} \right) \tag{3.2}$$

  Where A indicates the attribute (or dependent event) of the active target, B indicates the attribute (or prior event) of the active predictor, $\rho(A)$ indicates the $'A'$ prior probability, $\rho(A|B)$ indicates the $'B'$ posterior probability, and $\rho(B|A)$ indicates the $'B'$ probability when $'A'$ hypothesis holds true.

3.2 Deep Learning Approaches

DL represents a category of ML models in which the classification is carried out via data training using multiple layers within hierarchical networks based on unsupervised learning. These networks are inventive from the brain's architectural depth. According to the University of Toronto, Deep Belief Network (DBN) is coming up, and the data is trained using an approach that greedily trains layer by layer based on unsupervised learning for every Restricted Boltzmann Machine (RBM) layer [33]. According to the finding by the researchers Hinton et al. [19] and utilizing the same concept, various deep networks are presented that provide an effective classification task. This new trend DL was utilized for cost prediction or estimation.

4 Evaluation Techniques

In this research, the best prediction tools were identified for regression and classification tasks. That can be confusion metrics, accuracy, and loss log for classification, while predictor for regression task can be measured via errors as a mean absolute percentage (MAP), mean squared (MSE), root mean squared (RMSE), mean absolute deviation (MAD), and the coefficient of determination (R^2) or adjusted R^2. These measurements are given as follows:

- **MSPE:** It indicates the average of the difference square between the original values and the predicted values [9]:

$$\left(\text{MSE} = \sum_{i=1}^{N} (\text{predicted cost} - \text{actual cost})\right) \tag{4.1}$$

- **MSAP:** It is the average absolute percentage difference between the labels values and the predicted values and comparing the estimated and actual outcomes [9]:

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - m(x)| \tag{4.2}$$

  Where, $m(x)$ =average value of the data set, n = number of data values, and x_i =data values in the set.

- **R square** (R^2)**:** R^2 coefficient of determination, SSE is the sum of squares of the residuals and SST is the total sum of squares is expressed as follows [13]:

$$\left(R^2 = 1 - \left(\frac{SSE}{SST}\right)\right) \tag{4.3}$$

- **Adjusted R Square:** R^2 represents an adjusting for the number of variables involved in the models where R^{*2} is lower than the R^2 value. This measure is ranged between 0 and 1; a higher value denotes a higher quality model. The error can be classified as an excellent estimation if MAPE is less than 10%, between 10% from 20% is a good estimation. Between 20% and 50% is acceptable forecasting and more than 50% is an inaccurate estimation [13]. Error % categorization:

  Below 10, $10\% \geq \text{Error} \geq 0$ Below 20, $20\% \geq \text{Error} > 10\%$ Unacceptable, $\text{Error} > 20\%$

  R^2 is computed as follows [13]:

$$\left(R^{*2} = R^2 - \frac{(1 - R^2)K}{n - (K + 1)}\right) \tag{4.4}$$

- **Precision:** Number of items correctly identified as positive out of total true positives [43]:

$$\left(Precision = \left(\frac{TP}{TP + FP}\right)\right) \tag{4.5}$$

  Where TP indicates True Positives which is the number of projects correctly predicted to cost, TN indicates True Negatives which is the number of projects that were not correctly costed and classified as Incorrect, FP indicates False is the Positives Number of projects whose cost was not correctly predicted and classified as correct, and FN indicates False Negatives which is the number of projects whose cost was not correctly forecast and classified as incorrect

- **Recall:** measures the number of correct classifications penalized by the number of missed entries identified as in [43]:

$$\left(Recall = \left(\frac{TP}{TP + FN}\right)\right) \tag{4.6}$$

- **Accuracy:** It indicates the ratio of the correct predictions for both TP and TN compared with the entire number of tested cases [43]:

$$\left(Accuracy = \left(\frac{TP + TN}{TP + TN + FN + FP}\right)\right) \tag{4.7}$$

- **Log Loss:** It is an accuracy measurement that utilizes the probabilistic confidence idea, as given by the following expression regarding binary class:

$$(Loss = ((y \log(p) + (1 - y) \log(1 - p)))) \tag{4.8}$$

Taking into consideration the unpredictability depends on how much it varies from the actual label.

5 Analysis and Discussion

Civil engineering is an essential area to keep civilization with development, the cost could be sensitive information in the various civil engineering research and making the decision to select offers for execution with predicted cost. The studies had been proposed to construct models using different approaches static or dynamic for estimating the cost based on considerable datasets or features of the project. Recently, the ML technique played an important in the dynamic approach for estimation or prediction model, these techniques are unequal performance for calculating the cost of the civil domain. The evaluating techniques depend on the type of project and effectiveness measurement. The quantitative measurement of ML, including such as errors, accuracy, precision, and recall rates. Which are used to check the ability of ML models as a predictor in civil research. Table 2 indicates recent research on civil engineering with a duration of 2017 to 2021. The quantitative measurement of ML, including such as errors, accuracy, precision, and recall rates, are used to check the ability of ML models as a predictor in civil research. Different countries are summed the construction dataset for various domains of construction, Figure 3 indicates unequal years' percentage of collection data sets, building field dataset is collected with the duration is started 1995 to 2019, while Roads dataset are collected started from 1993 to the 2017 year, the time collection of the bridge from 2005 to 2015 as shown later in Table 2.

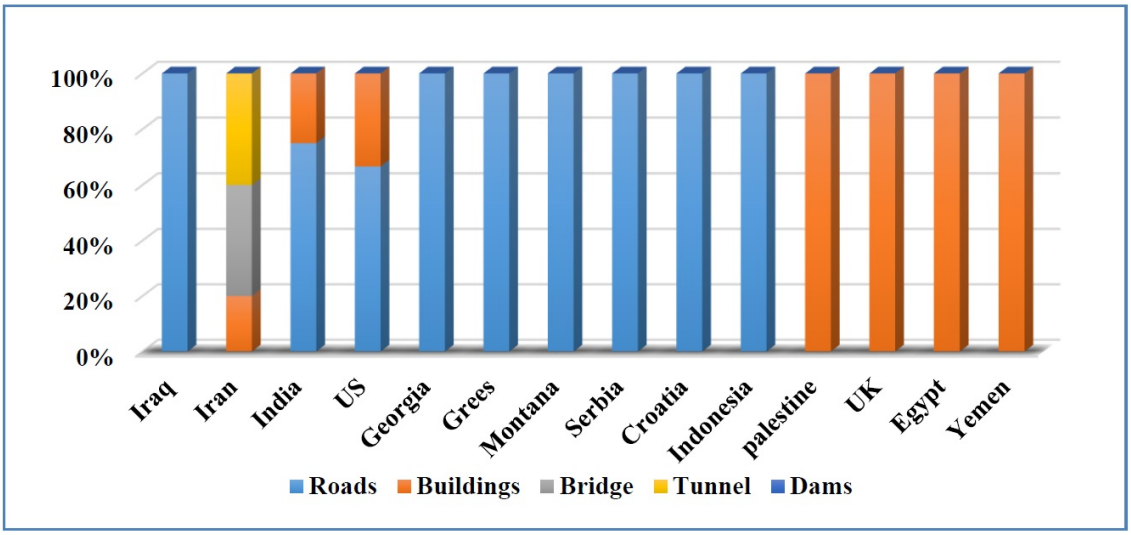

Figure 3: Project's country.

Figure 4 depicts the algorithms of ML and DL models that were used in each field of construction domains. It is clear that the ANN has been receiving much more attention compared with other algorithms, ANN is almost observed in each cost estimation for construction domains. In the construction cost prediction, most ML models were used as models to come rank first ANN than algorithms in different fields. The building cost prediction is calculated by NN, SVM, DT, RF, and NB models with 5%, 3%, 4%, 2%, and 3%, respectively. While the roads cost estimation is clearly used the ANN a lot with the percentage of 12% and SVM is the second one with 5% and Just like bridges cost estimation using ANN and SVM are used with a slight percentage about 3% and 2% respectively. It is obvious,

that the costs of tunnels and dams are calculated with a percentage of 2% by using NB. The airports and railroads field seems like an unexploited area because the low number of papers on airports and dams may be the scarcity of airports and the railroad's construction compared with other fields of construction domains and the estimation cost for Airports and Railroads didn't use machine learning algorithms. It is clear and as demonstrated in Figure 5 based on Table 1 that the ANN dominates this field due to the high accuracy it provides in such models, where the accuracy of the results ranges between 100% and 90%. Accuracy measures have been relied upon in evaluating the performance of most predictive models in the various domains of constructions, where accuracy refers to how close the measurement is to the true or acceptable value.

Several evaluation Metrics are used for measuring the ability of ML/DL to predict the cost in different domains of construction, the error, accuracy, and R^2 are generally preferred for evaluating many domains in the construction projects as shown in Figure 4. Where the error rate is higher used as a measurement to evaluate the Road project and less used with Building projects, few studies are used the error measurement for cost predictors with Airport, Dams, and Tunnels projects. Accuracy measurement is proposed to evaluate accurately the ML approaches as predictor models, the Building studies are much-used accurate than other projects like Road, Airport, and Dam. As R^2 measurement is proposed to measure the ability of predictor system and with high percentage for Road projects, average usage for Building projects. The precision metrics is much used with Road project with few studies that are used in other projects.

6 Current Issues and Recommendation

Basic on the actual cost, which is become a high cost to construct any civil project, the AI approach proposed to estimate the cost for a construction project. That could be dangerous to calculate the actual cost if it is lost or error, it can be found as, an ML predictor. It needs to highlight weakness points found recently AI researches as follows:

- **Lack Dataset:** Regression models in AI, that can be predicted the cost is approximately equal to the cost of truth as possible based on the validation dataset. As shown in Figure 3, there are some lack case datasets, most of the presented dataset was inefficiently collected depending on the personality stations, and the most cost estimation dataset is made of a set of variables according to quantity research with few samples or instances. Furthermore, this review presented a few datasets that are created to construct AI models based on our research query. It is clear that the time period in which the data was used is outdated, as the construction field used data between 1993 to 2019. As for the roads field, the data covered the years between 1997 and 2019, also the Bridges only one data set between 2005 to 2018. Studies indicate that the data that was used to predict costs for the construction and roads field is outdated and the sample counters are too few for a prediction model in machine learning and deep learning. It was found during the research that the data that was worked on by researchers in all fields of construction and roads has not been published except for one set of data related to forecasting the cost of buildings in Iran has been published as seen in Table 2.

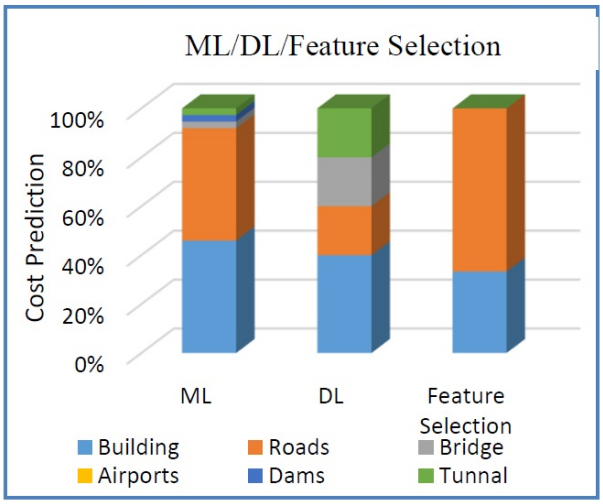

Figure 4: ML and DL approach studies.

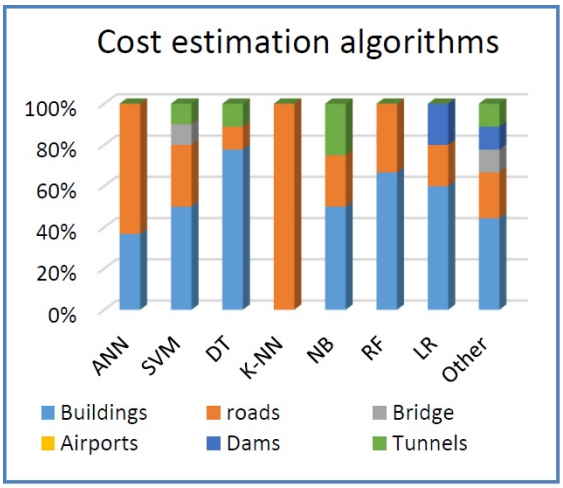

Figure 5: Prediction metrics

- **Impractical task for a cost:** The classification is used with ML approach than regression task. Unfortunately, the construction domain is different from other domains like images, networks that required prediction cost with linear prediction. The Methodology of different AI methods, which are conducted in isolation, do not measure fair in terms of effectiveness and efficiency, as shown in Table 2. This is due to diversity in the used dataset, pre-processing, and the configuration of ML/DL, and H/W platforms. So, there is a requirement for more benchmarking experiment studies that utilize a platform of unified computing and common impacting factors of various architectures for obtaining a fair outcome.
- **Invalidation Methodology:** The Methodology of different AI methods, which are conducted in isolation, do not measure fair in terms of effectiveness and efficiency, as shown in Table 2. This is due to diversity in (1) the used dataset, (2) pre-processing, (3) ML/ DL configuration, and (4) hardware platforms. Therefore, there is a need for more benchmarking experimental studies that use a unified computing platform and common affecting factors for different AI architectures in order to obtain a fair result.

We recommend the publication of ML- cost estimation-based data sets. Only publicly available data sets can be used by third parties and thus serve as a basis for evaluating cost prediction. Likewise, the quality of data sets can only be checked by third parties if they are publicly available. Last but not least, we recommend the publication of additional metadata such that third parties are able to analyze the data and their results in more detail most plans of the government are designed and proposed with at least five-year plans. Furthermore, this area will be more general and public for interesting researchers, it is a requirement the central for the data collection, analysis units by the government.

Table 2: The Methodology of different AI-methods

Ref.	Cost Estimation Project	Dataset Location	Dataset Collection	ML	DL	Effectiveness		
						MAPE	Accuracy	R^2
[1]	highway	Iraq	✗	ANN	✗	6.81	93%	0.8105
[16]	Bridges	Canada	✗	Genetic algorithm	✗	✗	95%	✗
[35]	Roads	Serbia	2005-2012 (166)	(ANNs) (SVM)	✗	0.0706, 0.2277	✗	✗
[2]	Building	India	2005-2016 (813)	OLS	✗	2.0217, 9.6617	91%, 97%	0.7122, 0.9923
[44]	roadway	U.S.	✗	maximum likelihood	✗	✗	✗	✗
[38]	Building	Iran	1993-2005 (372)	SVM and BPNN	DBM	✗	90.3%	✗
[27]	Building	Poland	2015 (143)	ANN	✗	0.03086	✗	0.84135
[6]	highway	Brazil	2010 to 2016 (14)	ANN	✗	0.008	99%	✗

[10]	highway	Georgia	2008 to 2016 (1400)	Gbm, Xgb, Rngr, ANN	✗	7.56	✗	0.47
[25]	Building	Poland	✗	ANN	✗	0.997	✗	✗
[11]	Building	Egypt	(51)	DT and NB	✗	0.488	51.2%	✗
[18]	Building	Yemen	2011-2015 (136)	ANN	✗	0.0014	99.86%	✗
[28]	Highway	India	2014–2017 (52)	ANN	✗	0.0846	91.54%	0.8960
[7]	Building	UK	(437, 000)	DT, RF, GBM	DNN	✗	80%	0.9672
[36]	Dams	Australia	(98)	Linear Regression	✗	0.2879	96%	0.5
[3]	Highway	Greece	1997-2015 (20)	FANN	✗	9.36965* 10–5	✗	✗
[24]	Bridge	Poland	2005-2018	SVM	✗	0.1094	✗	0.98
[11]	Building	Egypt	2002-2018	ANN, LR	✗	8.3	✗	✗
[30]	Tunnel	Iran	✗	GPR, SVR, DT	✗	0.0018	✗	0.97
[12]	Building	Egypt	✗	MLs	✗	0.09091	✗	0.92
[47]	Roads	Croatia	1999-2019	ANN	✗	0.1306	✗	0.9595
[26]	Highway	Montana	2006 and 2015 (996)	MRA, ANN	✗	0.25	✗	0.73
[45]	Building	U.S.	2003 - 2019 139	Mls	✗	0.1820	✗	✗
[22]	Building	India	2005-2016 813	MLs	✗	9.6617	97%	0.9923
[42]	Road	India	124	ANN	DNN	0.703	✗	✗
[40]	Road	Indonesia	2012-2017	✗	RBM	0.586	✗	✗
[49]	Road	India	2000-2018 363	ANN	✗	0.0502	✗	0.824
[29]	Roads	U.S.	2001 – 2017 (14,076)	MLs	✗	✗	92.51%	✗
[41]	Buildings	UK	60,000	DT, MDT, LightGBM, XG-Boost	✗	7.28	92%	✗
[21]	Building	Palestine	46	LR	✗	0.05	✗	96% 95%
[31]	Tunnel	Iran	✗	✗	RBM	✗	✗	✗

7 Conclusion

Most of the construction cost dataset is more collecting and used for training and testing ML models in buildings and road projects, while few other fields of ML studies are concerned with collecting datasets in Dams, Tunnels, and Airports. Most existing studies have proven the possibility of building a high-performing model of ML for predicting constructions cost. The ML models are more effective than other conventional simulation models since they are capable of returning cost outcomes in seconds. However, there is no identified optimal model of ML to achieve the cost prediction. Unfortunately, few DL approaches were proposed for predicting the cost of estimation. Most ML models utilized in cost prediction such as SVM, DT, RF, NB, K-NN, ANN, and others. Typically, SVM and ANN have produced the best outcomes and exceeded other models for building and road projects. In future works, the regression task is more required for cost prediction using deep learning approaches to be a more effective model.

References

[1] F.M.S. AL-Zwainy and I.A.-A. Aidan, *Forecasting the cost of structure of infrastructure projects utilizing artificial neural network model (highway projects as case study)*, Indian J. Sci. Technol. **10** (2017), no. 20, 1–12.

[2] S.S. Arage and N.V. Dharwadkar, *Cost estimation of civil construction projects using machine learning paradigm*, Proc. Int. Conf. IoT in Social, Mobile, Analytics and Cloud, I-SMAC **2017** (2017), 594–599.

[3] G.N. Aretoulis, *Neural network models for actual cost prediction in Greek public highway projects*, Int. J. Proj. Organ. Manag. **11** (2019), no. 1, 41.

[4] M.I.N.S.O.O. Baek and B.A.A.B.A.K. Ashuri, *Spatial regression analysis for modeling the spatial variation in highway construction costs, Resilient Structures and Sustainable Construction*, Georgia Institute of Technology, Atlanta, 2017.

[5] M. Barakchi, O. Torp and A.M. Belay, *Cost estimation methods for transport infrastructure: a systematic literature review*, Proc. Eng. **196** (2017), 270–277.

[6] L.B. Barros, M. Marcy and M.T.M. Carvalho, *Construction cost estimation of Brazilian highways using artificial neural networks*, Int. J. Struct. Civ. Eng. Res. **7** (2018), no. 3, 283–289.

[7] M. Bilal and L.O. Oyedele, *Guidelines for applied machine learning in construction industry–a case of profit margins estimation*, Adv. Engin. Inf. **43** (2020), 101013.

[8] Bureau of Infrastructure, *Transport and regional economics (BITRE), road construction cost and infrastructure procurement benchmarking: 2017 update*, BITRE, Canberra ACT, 2018.

[9] Cambridge Dictionary, *Meaning of airport in English*, https://dictionary.cambridge.org/dictionary/english/airport, 2022.

[10] Y. Cao, B. Ashuri and M. Baek, *Prediction of unit price bids of resurfacing highway projects through ensemble machine learning*, J. Comput. Civ. Eng. **32** (2018), no. 5, 04018043.

[11] Y. Elfahham, *Estimation and prediction of construction cost index using neural networks, time series, and regression*, Alexandria Eng. J. **58** (2019), no. 2, 499–506.

[12] H.H. Elmousalami, *Artificial intelligence and parametric construction cost estimate modeling: state-of-the-art review*, J. Constr. Eng. Manag. **146** (2020), no. 1, 03119008.

[13] H.H. Elmousalami, *Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis*, IEEE Trans. Engin. Manag. **68** (2021), no. 1, 183–196.

[14] M. Flah, I. Nunez, W. Ben Chaabene and M.L. Nehdi, *Machine learning algorithms in civil structural health monitoring: a systematic review*, Arch. Comput. Method in Engin. **28** (2021), no. 4, 2621–2643.

[15] D.M. Frangopol, Y. Dong and S. Sabatino, *Bridge life-cycle performance and cost: analysis, prediction, optimisation and decision-making*, Struct. Infrastruct. Eng. **13** (2017), no. 10, 1239–1257.

[16] F. Ghodoosi, S. Abu-Samra, M. Zeynalian and T. Zayed, *Maintenance cost optimization for bridge structures using system reliability analysis and genetic algorithms*, J. Constr. Eng. Manag. **144** (2018), no. 2, 04017116.

[17] A. Gondia, A. Siam, W. El-Dakhakhni and A.H. Nassar, *Machine learning algorithms for construction projects delay risk prediction*, J. Constr. Eng. Manag. **146** (2020), no. 1, p. 04019085.

[18] W. Hakami and A. Hassan, *Preliminary construction cost estimate in Yemen by artificial neural network*, Balt. J. Real Estate Econ. Constr. Manag. **7** (2019), no. 1, 110–122.

[19] G.E. Hinton, S. Osindero and Y.-W. Teh, *A fast learning algorithm for deep belief nets*, Neural Comput. **18** (2006), no. 7, 1527–1554.

[20] M. Hu and M.J. Skibniewski, *A review of building construction cost research: current status, gaps and green buildings*, Green Build. Constr. Econ. *2* (2021), no. 1, 1–17.

[21] A. Issa, R. Bdair and S. Abu-Eisheh, *Assessment of compliance to planned cost and time for implemented municipal roads projects in Palestine*, Ain Shams Eng. J. **13** (2022), no. 2, 101578.

[22] A. Jaafari, I. Pazhouhan and P. Bettinger, *Machine learning modeling of forest road construction costs*, Forests **146** (2021), no. 9, 1169.

[23] M. Juszczyk, *The challenges of nonparametric cost estimation of construction works with the use of artificial intelligence tools*, Proc. Eng. **196** (2017), 415–422.

[24] M. Juszczyk, *On the search of models for early cost estimates of bridges: an SVM-based approach*, Build. **10** (2020), no. 1.

[25] M. Juszczyk, A. Leśniak and K. Zima, *ANN based approach for estimation of construction costs of sports fields*, Complexity **2018** (2018).

[26] I. Karaca, D.D. Gransberg and H.D. Jeong, I*mproving the accuracy of early cost estimates on transportation infrastructure projects*, J. Manag. Eng. **36** (2020), no. 5, p. 04020063.

[27] A. Leśniak and M. Juszczyk, *Prediction of site overhead costs with the use of artificial neural network based model*, Arch. Civ. Mech. Eng. **18** (2018), no. 3, 973–982.

[28] G. Mahalakshmi and C. Rajasekaran, *Early cost estimation of highway projects in India using artificial neural network*, Lecture Notes in Civil Engineering, 25 (2019), 659–672.

[29] A. Mahdavian, A. Shojaei, M. Salem, J.S. Yuan and A.A. Oloufa, *Data-driven predictive modeling of highway construction cost items*, J. Constr. Eng. Manag. **147** (2021), no. 3, 04020180.

[30] A. Mahmoodzadeh, M. Mohammadi, A. Daraei, H. Farid Hama Ali, A. Ismail Abdullah and N. Kameran Al-Salihi, *Forecasting tunnel geology, construction time and costs using machine learning methods*, Neural Comput. Appl. **33** (2021), no. 1, 321–348.

[31] A. Mahmoodzadeh, M. Mohammadi, S. Nariman Abdulhamid, H. Hashim Ibrahim, H. Farid Hama Ali and S. Ghafoor Salim, *Dynamic reduction of time and cost uncertainties in tunneling projects*, Tunnel. Underground Space Technol. **109** (2021).

[32] Z.K. Maseer, R. Yusof, N. Bahaman, S.A. Mostafa and C.F.M. Foozy, *Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset*, IEEE Access **9** (2021), 22351–22370.

[33] J.D. McCaffrey, *Restricted boltzmann machines are difficult to explain*, Software Research, Development, Testing, and Education, https://jamesmccaffrey.wordpress.com/, (2016).

[34] F. Ning, Y. Shi, M. Cai, W. Xu and X. Zhang, *Manufacturing cost estimation based on a deep-learning method*, J. Manufactur. Syst. **54** (2020), 186–195.

[35] I. Peško, V. Mučenski, M. Šešlija, N. Radović, A. Vujkov, D. Bibić and M. Krklješ, *Estimation of costs and durations of construction of urban roads using ANN and SVM*, Complexity **2017** (2017).

[36] C. Petheram and T.A. McMahon, *Dams, dam costs and damnable cost overruns*, J. Hydrol. X **3** (2019).

[37] F. Pour Rahimian, S. Seyedzadeh, S. Oliver, S. Rodriguez and N. Dawood, *On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning*, Autom. Constr. **110** (2020).

[38] M.H. Rafiei and H. Adeli, *Novel machine-learning model for estimating construction costs considering economic variables and indexes*, J. Constr. Eng. Manag. **144** (2018), no. 12, p. 04018106.

[39] K. Riga, K. Jahr, C. Thielen and A. Borrmann, *Mixed integer programming for dynamic tower crane and storage area optimization on construction sites*, Autom. Constr. **120** (2020).

[40] A. Sazali, B.H. Setiadji and B. Haryadi, *Prediction of road handling cost using Markov chain method in regency road network*, Int. J. Integr. Eng. **13** (2021), no. 4, 275–283.

[41] A. Shehadeh, O. Alshboul, R.E. Al Mamlook and O. Hamedat, *Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree*, LightGBM, and XGBoost regression, Autom. Constr. **129** (2021).

[42] N. Suneja, J.P. Shah, Z.H. Shah and M.S. Holia, *A neural network approach to design reality oriented cost estimate model for infrastructure projects*, Reliab. Theory Appl. **16** (2021), 254–263.

[43] A. Swalin, *Choosing the right metric for evaluating machine learning models–part 2*, Data Institute, 2018.

[44] O. Swei, J. Gregory and R. Kirchain, *Construction cost estimation: a parametric approach for better estimates of expected cost and variation*, Transport. Res. Part B: Methodol. **101** (2017), 295–305.

[45] N. Tajziyehchi, M. Moshirpour, G. Jergeas and F. Sadeghpour, *A predictive model of cost growth in construction projects using feature selection*, IEEE Third Int. Conf. Artificial Intell. Knowledge Engin. (AIKE), IEEE, 2020, pp. 142–147.

[46] S. Tayefeh Hashemi, O.M. Ebadati and H. Kaur, *Cost estimation and prediction in construction projects: a systematic review on machine learning techniques*, SN Appl. Sci. **2** (2020), no. 10.

[47] K. Tijanić, D. Car-Pušić and M. Šperac, *Cost estimation in road construction using artificial neural network*,

Neural Comput. Appl. **32** (2020), no. 13, 9343–9355.

[48] T.N. Van and T.N. Quoc, *Research trends on machine learning in construction management: a scientometric analysis*, J. Appl. Sci. Tech. Trends **2** (2021), no. 3, 96–104.

[49] P. Velumani, N.V.N. Nampoothiri and U. Mariusz, *A comparative study of models for the construction duration prediction in highway road projects of India*, Sustain. **13** (2021), no. 8, 4552.

[50] J.T. Von Brown, *A planning methodology for railway construction cost estimation in North America*, Iowa State University, 2011.

[51] Y. Xu, Y. Zhou, P. Sekula and L. Ding, *Machine learning in construction: from shallow to deep learning*, Dev. Built Envir. **6** (2021), 100045.