IOP Conference Series:
Earth and
Environmental Science

PAPER • **OPEN ACCESS**

Comparing two AI methods for predicting the future trend of New Zealand building projects: Decision Tree and Artificial Neural Network

To cite this article: A Zavvari *et al* 2022 *IOP Conf. Ser.: Earth Environ. Sci.* **1101** 082016

Comparing two AI methods for predicting the future trend of New Zealand building projects: Decision Tree and Artificial Neural Network

A Zavvari[1], M B Jelodar[1] and M Sutrisna[1]
[1] School of Built Environment, Massey University, Auckland 0745, New Zealand

a.zavvari@massey.ac.nz

Abstract. The rise of Artificial Intelligence and Machine Learning in many aspects of construction management has helped this industry to further improve the management, design, and planning of construction projects. This trend happens in many construction sectors, including in New Zealand. Whilst relatively smaller compared to construction sectors in other OECD countries, the construction sector in New Zealand carries a similar degree of complexity and with its own unique characteristics. Various studies showed that AI and ML can be used for analysis of construction data to generate further insights and to predict future trends in construction sectors. However, the AI approaches have their own set of challenges such as complexity, high cost of training, failure, and change. Aiming to better understand the trends and requirements of New Zealand building projects, this study started with a review of the existing AI methods that are currently being applied. Accordingly, compare and evaluate the accuracy of two AI prediction methods. The two methods of Decision Tree and Artificial Neural Network are selected based on their predictive power and accuracy. These methods are conducted by using available historical building data which is available on StatsNZ website. A portion of the data is used for testing and evaluation purposes, and the rest of the data is used for training the AI methods. It was identified that the Decision Tree method did not show suitable accuracy for prediction building consents issued data. In comparison, Artificial Neural Network shows a reasonable range with 95% of confidence level. Therefore, this method is applied for building consents issued in New Zealand.

1. Introduction

Today's world gets overwhelmed with data, and modern technologies are steadily increasing. Many industries challenge is how to deal with a massive volume of their daily data [1, 2]. Data is a collection of facts that can be collected as text, numbers, images, videos, web pages, measurement, description, or observation. There are different types of data structures: structured, unstructured, and semi-structured. Structured data is the highly organised and formatted data ready for integration with Artificial Intelligence (AI) technologies. On the other hand, unstructured data is unformatted and not organized in a predefined way. Semi structured data is a combination of structured and unstructured data [3].

During the past decade, technologies and digital construction have become a new way for the construction industry [4]. Nowadays, a massive number of modern technologies such as drones, AI cameras, robotics arms, building information modeling (BIM), virtual reality, augmented reality, internet of things, impacting the construction industry, are introduced to construction [5-8]. These new technologies generate a significant volume of data every day. For example, a cloud-based BIM collects life cycle data of each construction project and allows project data to be accessible in any location and

at any time [9]. The data collection started from planning to conceptual design to detailed design, which included all of the changes, analysis, and communications in the design team. Then, it gets documented and carried to the fabrication team. All the data during the construction, operation, maintenance, renovation, or demolition are collected and saved in the system. These new technologies produce a different type of data with various types of formatting filling up the databases. This unstructured data makes the data analysis difficult, especially at the government level of decision making [10]. On the other hand, the construction industry deals with a significant volume of data, which allows scientists to find valuable information and knowledge. The rapidly developing technologies gave these opportunities to every industry to have powerful and high-speed internet and devices. The devices get connected to the internet and integrate into other devices, and it's the base stage of the internet of things [11].

As the data grows, Machine Learning (ML) algorithms as a form of AI are able to make the prediction based on the collected data. Having access to this real-time data and using ML algorithms make project management and decision-making easier than before. Lots of ML algorithms have already been introduced to the construction, such as Support Vector Machine, Artificial Neural Network, Decision Tree, k-Nearest Neighbours, Random Forest, Naive Bayes [7].

Although the mentioned methods are applied in different construction areas, few studies have been conducted for building consents issued data. Therefore, this study applied common AI methods as Artificial Neural Network and Decision Tree to investigate the best compatible and suitable AI method for building consents issued data which is publicly available on StatNZ website.

2. Literature Review

This research focuses on the Artificial Neural Network (ANN) and Decision Tree (DT) algorithm as the best-performed algorithm in construction data. The ANN is one of the most used methods in recent years, which is defined as a reasoning-based model of the human brain. The human brain consists of neurons, while synapses create the connections between the neurons. By using multiple neurons simultaneously, the brain can perform its functions. Each neuron has a very simple structure, but a large number of such elements constitute a tremendous processing power. The functions of the natural neurons are stimulated by the artificial neurons. Artificial neurons receive a number of inputs and based on defined functions, calculate the outputs. Another common prediction method is DT. The DT as a fast algorithm involves root-proposing, intermediate, and leaf nodes. Each tree node is a decision path to an outcome. The DT approach partitions a dataset input space into reciprocally excluded domains, labeling each with values or actions describing their data points. Splitting criteria are used to determine the attribute which best splits that partial tree of the training data which attains a specific node.

2.1. Decision Tree

A Decision Tree is a common and fast classification/prediction method. DTs have been successfully applied in various fields of science as well as construction. DTs have different approaches, such as Random Forest (RF) and Decision Tree J48 (DT-J48). The general DT and mentioned approaches are explained as follows.

DTs classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance. Information gain of the DT is based on the concept of entropy from information theory. Entropy is a measure of the uncertainty associated with a discrete random variable, where C is the set of desired classes. A set of examples D is possible to compute the original entropy of the D dataset as given by equation (1).

$$\mathrm{H(D)} = -\Sigma \mathrm{P(c}j\mathrm{)log2P(c}j\mathrm{)\ C}j = 1 \tag{1}$$

Entropy of an attribute Ai with v values will partition D into v subsets. The expected entropy of Ai is calculated by equation (2).

$$\mathrm{HA}i\mathrm{(D)} = \Sigma|\mathrm{D}j\ ||\mathrm{D}|\ \mathrm{H(D}jvj = 1) \tag{2}$$

In general, the information gained by selecting attribute Ai to branch or to partition the data is given by the difference of prior entropy and the entropy of the selected branch (see equation 3).

$$\text{GAIN}\,(\text{D}, \text{A}i) \;=\; \text{H}(\text{D}) - \;\text{HA}i(\text{D}) \tag{3}$$

The Random Forest (RF) method is an approach of DT which uses a collection of de-correlated or independent trees [12]. Each tree generates a class prediction for a given observation, and a majority voting scheme is used to generate the predicted class for the ensemble. To form the de-correlated trees, each tree is grown on an independent bootstrap sample, and at each node, *m* variables are selected randomly out of all *m* possible input variables to determine the best split at that node. The best split from the *m* variables is then selected. Different measures of node impurity can be used to select the best split at each node of the tree. Three types of node impurity measures are misclassification rate, Gini index, and cross-entropy or deviance (see equation 4 - 6). Where k (*m*) is the majority class in node *m*; P*mk* is the proportion of class *k* observations in node *m*; y*i* is the class for observation *i*. The Gini index measures how often a randomly chosen observation from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. The Gini index and cross-entropy measures are more sensitive to changes in node probabilities than the misclassification rate [13, 14].

$$\text{Misclassification rate} = \;1\;\text{N}m\;\Sigma\text{I}\;(\text{y}i \neq k(m)) = 1 - \text{P}mki \in \text{R}m \tag{4}$$

$$\text{Gini index} = \;\Sigma\text{P}mk(1 - \text{P}mk)\text{K}k = 1 \tag{5}$$

$$\text{Cross} - \text{entropy or deviance} = -\Sigma\text{P}mk\;\text{logP}mk)\text{K}k = 1 \tag{6}$$

More recently, the DT approach with its rapid algorithmic output is being further applied in construction. The Design tree algorithms are applied in different aspects of construction. This algorithm is used as a powerful classifier for vehicle identification [15]. Data was collected by laser distance measuring and radar sensors. Another area of applying DT was to predict rock underground excavation and evaluate the level of rockburst based on microseismic monitoring data [16]. DT also show a good performance in predicting the sick building syndrome for health and lifestyle habit [17]. Currently, DT is a viral method, and the accuracy of the method is directly related to the quality of the input data. DT is a part of predicting wastewater and heavy construction [18, 19]

2.2. Artificial Neural Network

The ANN can be defined as a model of reasoning based on the human brain. The ANN consists of basic information-processing units called artificial neurons. Artificial neurons receive a number of inputs, and each input has its own weight. The input is multiplied with its own weight, and it sums with an additional bias number. After this process, it goes through a function process. In the end, the result is shown as an output. Figure 1 illustrates the information process in an artificial neuron [20, 21].

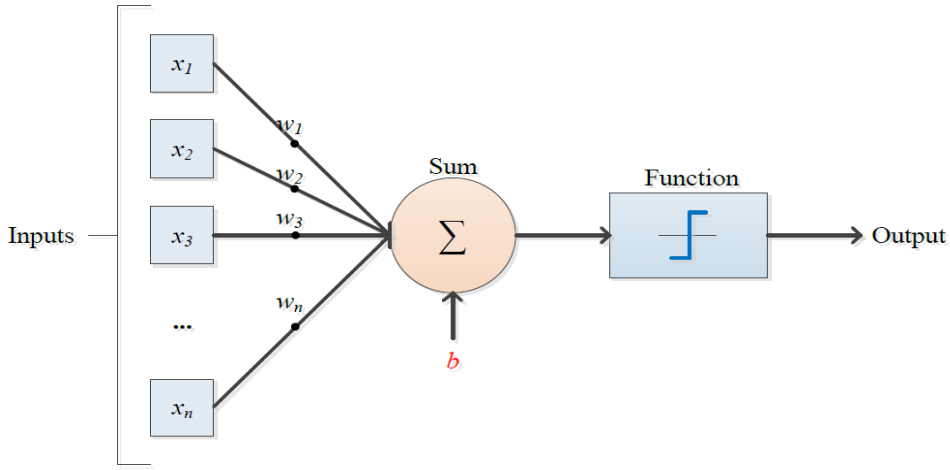

Figure 1. Information processing with an artificial neuron

ANN is one of the most powerful approximation multidimensional functions. It provides an exciting alternative method for solving a variety of problems in different fields of science and engineering. It has very wide applications in machine learning such as prediction, pattern recognition, function approximation, optimization and signal de-noising. ANN method is a set of connected input/output units where each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct classification output of the input signals. In this algorithm, a number of neurons work together and create a network. In general, artificial neurons receive a number of inputs as $x1, x2 \dots xi$, and each of them multiplies its own weight $w1, w2 \dots wi$; it sums with an additional bias b number. After that, it goes through function φ [22]. Equation (7) shows the neural network formula in general.

$$Y = \varphi\ (w1 \times x1 + w2 \times x2 + \cdots +\ wi \times xi +\ b),\ Y = \varphi\Sigma(\ wi \times xi +\ b)ni = 1 \qquad (7)$$

The ANN algorithm is one of the popular artificial intelligence methods which are widely applied on construction. An ANN- based model is applied for cable trust structure prediction to predict the new cases [23]. And this algorithm shows high accuracy in comparison with other methods. In their recent paper, Liu and Iqbal used ANN methods to predict diffusivity in concrete [24].

3. Research Methodology

The individual prediction methods such as DT and ANN read the data from a dataset consisting of different attributes and class data (target value). The dataset is divided into train-set and test-set for the training of the algorithm and prediction purposes, respectively. The algorithm based on train-set attributes and class data starts to learn the data and produces a model based on the data. The classifier then applies the produced model into the test-set (without class data) and starts to predict the class data value. Finally, based on the correct prediction of the class value, the accuracy of the system is calculated. Figure 2 shows the general view of the artificial intelligence prediction (based on classification) process for an individual algorithm.

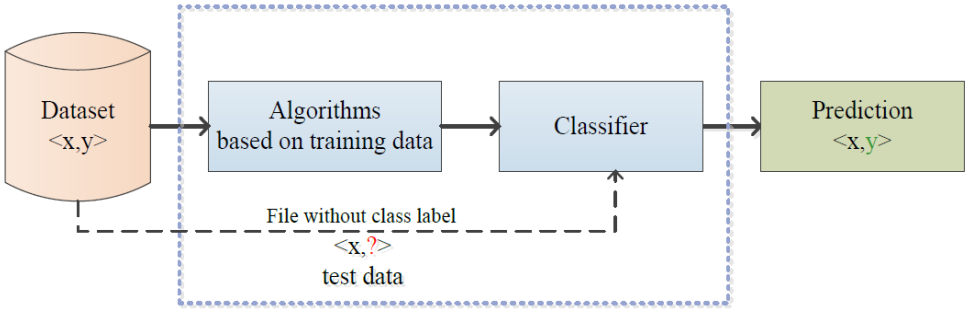

Figure 2. General views of individual algorithm classification/prediction

In recent years, to improve the accuracy of the prediction methods, the new approaches of applying multiple methods have been implemented in the area of the *k*-fold cross-validation evaluation model is applied to divide the dataset into the train-set and test-set. *k*-fold cross-validation is a systematic process of repeating the train-set and test-set split procedure multiple times in order to reduce the variance associated with a single trial of train-set and test-set split. This method evaluates and compares algorithms by dividing an entire dataset into two train-set and test-set. The general ideas of cross-validation are explained as follows (see figure 3).

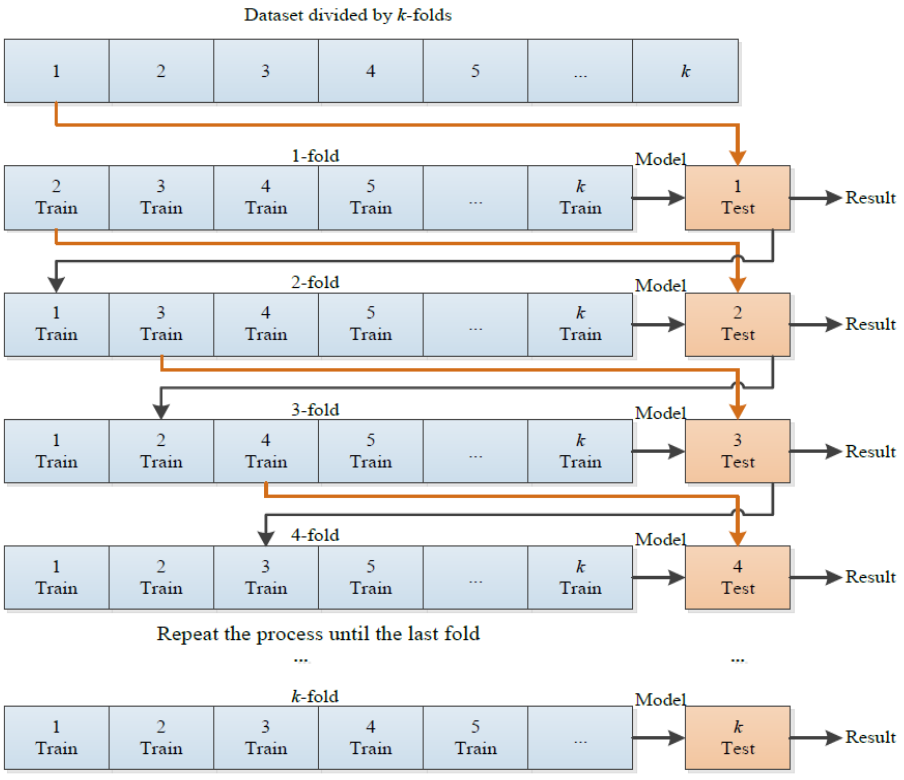

Figure 3. *k*-fold cross-validation structure

Cross-validation techniques divided the dataset into *k* partitions or folds (1, 2, 3, 4… *k*). In the first training time, the system uses folds (2, 3, 4… *k*) to train the model and tries to test the classifier based on the first fold. In the next step, folds (1, 3, 4… *k*) are used for training the model and tests the classifier on the left-out part. In each model *k-1* folds are used for training and one part is not used in training that model.

The model is tested on the fold that was not used in model training. In other words, the dataset divides the data into *k* parts, and then learns *k* models each time by leaving out one part of the training data, which is used for testing. The average of all results during *k*-fold cross-validation is the final result. The learning 10-fold cross-validation ($k = 10$) has the most common and reliable estimated accuracy A survey of cross-validation procedures for model selection [25, 26]. The main goal of cross-validation is comparing the performance of two or more different algorithms and finding out the best algorithm for the available data. Figure 4 shows 10-fold cross-validation applying for an individual method.

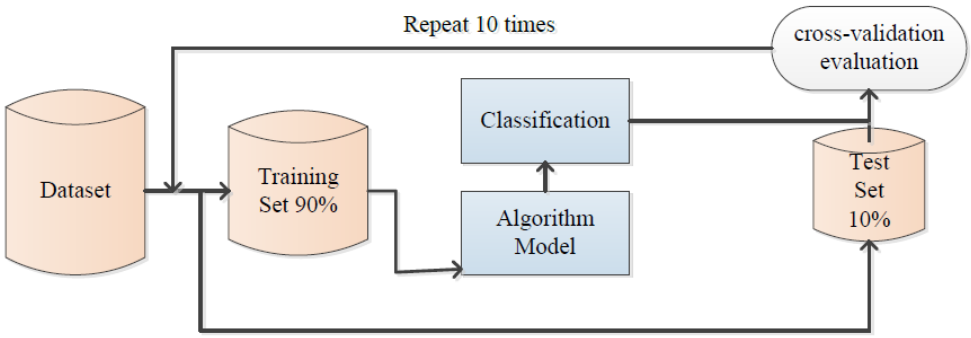

Figure 4. *10*-fold Cross-validation evaluation for individual model

4. Findings and Discussion

This paper used the residential building consents data is available at *https://www.stats.govt.nz/information-releases/* website page from Jan-2016 till the end of Dec-2021. Data is released annually on the website. The dataset includes the number of Houses, Apartments, Retirement village units, Townhouses, flats, and units, all dwellings, as shown in figure 5.

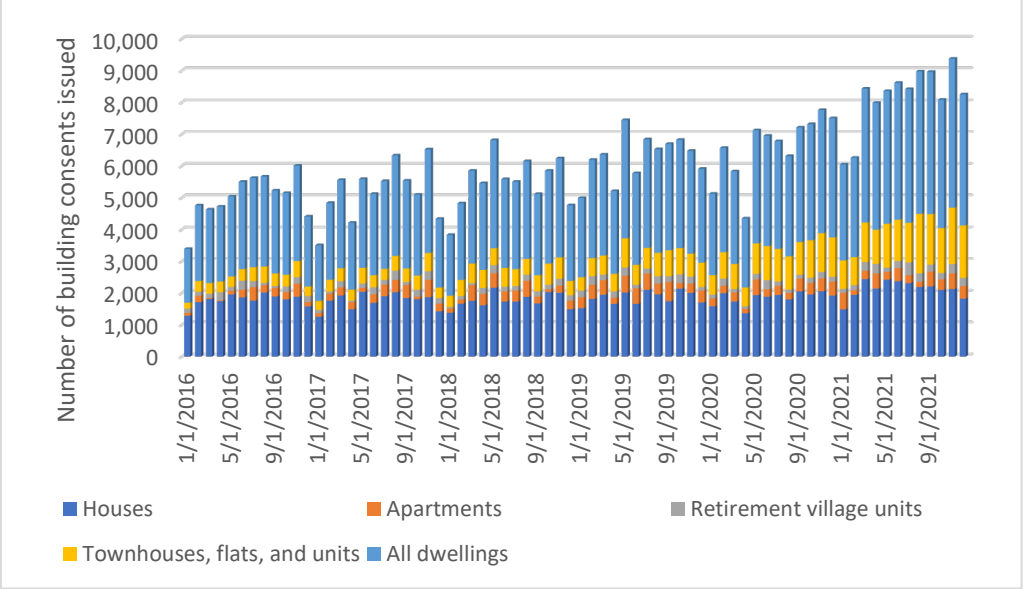

Figure 5. Number of building consents issued in 5 years (2016-2021)

Two AI methods (DT and ANN) are employed in the same dataset. The power BI platform is applied for visualization and Python as the programming language. In this case, 70% of data was applied for training the model, and 30% of data was treated for test-set and validation. The general explanation of the code in Python are explained as bellow (see figure 6):

```
1   # Load the dependencies
2   # Import Numpy and Pandas libraries
3   # From sklearm.model_selection import train_test_split
4   # From sklean import the models (ANN and DT)
5   # Load the database

6   Dataset = pd.read_csv (name of the dataset)
7   # Preprocess your data
8   # Split and train the dataset

9   X_train, x_test, y_train, y_test= train_test_split(x,y)
10   # Call the model to predict the result
```

Figure 6. General explanation (pseudo-code) for both algorithm

The power BI visualisation methods used for the ANN method show a better prediction result than DT. As illustrated in figure 7, Data are predicted for six months in advance with a 95 percent confidence level.

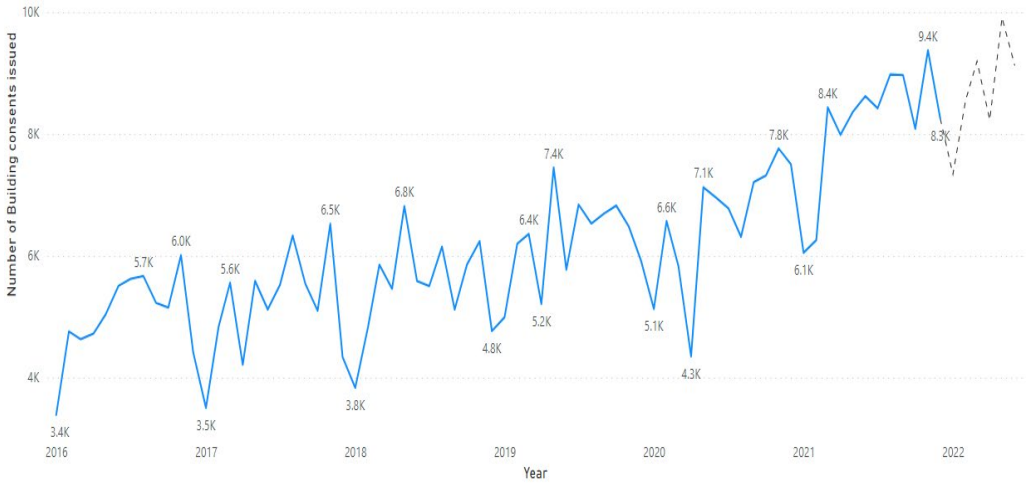

Figure 7. data prediction algorithm based on the number of building consents with 95% confidence level.

The dash-line shows the ANN prediction with 95% confidence level of the first six months of 2022 (Jan- 7338, Feb-8552, March-9196, April-8230, May-9911, Jun-9126). DT shows 50% accuracy for prediction, which is not enough power to predict the pattern of the building consents issued.

5. Conclusions

The current study aimed to determine the compatible methods for building consents issued in New Zealand based on publicly available data. DT and ANN as two AI methods were applied to predict the total number of building consents issued. The methods were programmed in Python and visualized in the Power BI desktop environment. The algorithms were validated based on 10-fold cross-validation. The 50 percent of accuracy resulted based on the DT as the first method, which was the insufficient level of accuracy. In comparison, ANN showed the result with a 95 percent confidence level. These experiments confirmed that the ANN algorithm achieved the highest accuracy based on the mentioned

dataset. The limitation of this study was the availability of data for five years from 2016 to 2021, which prevented a comprehensive overview of the pattern of data. The results of this research support the idea that AI methods can be applied in various construction fields. This new understanding should improve future forecasting of data in the construction industry. This study can be an initial step for predicting the future capability of this industry.

References

[1] Bilal M, Oyedele LO, Qadir J, Munir K, Ajayi SO, Akinade OO, et al. 2016 Big Data in the construction industry: A review of present status, opportunities, and future trends *Advanced engineering informatics* **30** 500-21
[2] Zavvari A, Sutrisna M and Jelodar MB 2022 *Evaluating Capacity and Capability of the Construction Sector: The Application of Big Data Tools* Published p. 457
[3] Martinez-Mosquera D, Navarrete Ra nd Lujan-Mora S 2020 Modeling and management big data in databases—A systematic literature review *Sustainability* **12** 634
[4] Eliwa H, Jelodar MB and Poshdar M *Information technology and New Zealand construction industry: An empirical study towards strategic alignment of project and organization* Published p. 21-3
[5] Chen H, Hou L, Zhang GK and Moon S 2021 Development of BIM, IoT and AR/VR technologies for fire safety and upskilling *Automation in Construction* **125** 103631
[6] Ham Y and Kamari M 2019 Automated content-based filtering for enhanced vision-based documentation in construction toward exploiting big visual data from drones *Automation in Construction* **105** 102831
[7] Pan Y and Zhang L 2021 Roles of artificial intelligence in construction engineering and management: A critical review and future trends *Automation in Construction* **122** 103517
[8] Eliwa HK, Jelodar MB and Poshdar M 2022 Information and Communication Technology (ICT) Utilization and Infrastructure Alignment in Construction Organizations *Buildings* **12** 281
[9] Babaeian Jelodar M and Shu F 2021 Innovative use of low-cost digitisation for smart information systems in construction projects *Buildings* **11** 270
[10] Cha H and Lee D 2018 Framework based on building information modelling for information management by linking construction documents to design objects *Journal of Asian Architecture and Building Engineering* **17** 329-36
[11] Wang H and Zheng BG *Research and Implementation of the Smart Home System Based on Internet of Things Environment* Trans Tech Publ Published p. 1915-8
[12] Breiman L 2001 Random forests *Machine learning* **45** 5-32
[13] Hastie T, Tibshirani R and Friedman J. Unsupervised learning. The elements of statistical learning (pp. 485-585). Springer, New York, NY; 2009.
[14] Strobl C, Malley J and Tutz G 2009 An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests *Psychological methods* **14** 323
[15] Wan HS, Jiang SM, Zhang Y and Zou GF *Vehicle identification and classification system based on Decision Tree* Trans Tech Publ Published p. 373-7
[16] Zhao H, Chen B and Zhu C 2021 Decision Tree Model for Rockburst Prediction Based on Microseismic Monitoring *Advances in Civil Engineering* **2021**
[17] Sarkhosh M and Najafpoor AA, Alidadi H, Shamsara J, Amiri H, Andrea T, et al. 2021 Indoor Air Quality associations with sick building syndrome: An application of decision tree technology *Building and Environment* **188** 107446
[18] Syachrani S, Jeong HSD and Chung CS 2013 Decision tree–based deterioration model for buried wastewater pipelines *Journal of Performance of Constructed Facilities* **27** 633-45
[19] Shehadeh A, Alshboul O, Al Mamlook RE and Hamedat O 2021 Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression *Automation in Construction* **129** 103827

[20] Makinde F, Ako C, Orodu O and Asuquo I 2012 Prediction of crude oil viscosity using feed-forward back-propagation neural network (FFBPNN) *Petroleum & Coal* **54** 120-31
[21] Krenker A, Bešter J and Kos A 2011 Introduction to the artificial neural networks *Artificial Neural Networks: Methodological Advances and Biomedical Applications InTech* 1-18
[22] Rojas R 2013 *Neural networks: a systematic introduction*: Springer Science & Business Media.
[23] Liu Z, Jiang A, Shao W, Zhang A and Du X 2021 Artificial-Neural-Network-Based Mechanical Simulation Prediction Method for Wheel-Spoke Cable Truss Construction *International Journal of Steel Structures* **21** 1032-52
[24] Liu Q-f, Iqbal MF, Yang J, Lu X-y, Zhang P and Rauf M 2021 Prediction of chloride diffusivity in concrete using artificial neural network: Modelling and performance evaluation *Construction and Building Materials* **268** 121082
[25] Karimi S, Yin J and Baum J 2015 Evaluation methods for statistically dependent text *Computational Linguistics* **41** 539-48
[26] Arlot S and Celisse A 2010 A survey of cross-validation procedures for model selection *Statistics surveys* **4** 40-79

Acknowledgment
The study is part of a research programme funded by the Ministry of Business Innovation and Employment (MBIE).