Automation in Construction 114 (2020) 103155

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring

Wei Dong[a], Yimiao Huang[a,*], Barry Lehane[a], Guowei Ma[a,b]

[a] Dept. of Civil, Environmental & Mining Engineering, The University of Western Australia, Perth 6009, Australia
[b] Dept. of Civil & Transportation Engineering, Hebei University of Technology, Tianjin 300401, China

ARTICLE INFO

Keywords:
Concrete electrical resistivity
Structural health monitoring
Machine learning
XGBoost algorithm

ABSTRACT

For structural health monitoring, electrical resistivity measurement (ERM) method is commonly employed for the detection of concrete's durability, as indicated by the chloride permeability and the corrosion of steel reinforcement. However, according to previous experimental studies, ERM results are susceptible to significant uncertainties due to multiple influencing factors such as concrete water/cement ratio and structure curing environment as well as their complex interrelationships. The present study therefore proposes an XGBoost algorithm-based prediction model which considers all potential influential factors simultaneously. A database containing 800 experimental instances composed of 16 input attributes is constructed according to existing reported studies and utilized for training and testing the XGBoost model. Statistical scores (RMSE, MAE and R^2) and the GridsearchCV feature are applied to evaluate and optimize the established model respectively. Results show that the proposed XGBoost model achieves satisfactory predictive performance as demonstrated by high coefficients of regression fitting lines (0.991 and 0.943) and comparatively low RMSE values (4.6 and 11.3 kΩ·cm) for both training and testing sets respectively. The analyses of the attribute importance ranking also reveal that curing age and cement content have the greatest influence on ERM results.

1. Introduction

Structural health monitoring (SHM) of civil infrastructure, such as buildings and bridges, has been employed for many decades. Appropriate SHM strategies are vital to ensure structures maintain their intended design safety factor under demanding working environments over the service lifecycle. In SHM, one of the critical issues is to assess the degree of degradation and the health status of structures including the chloride permeability and fractural damage [1]. A non-destructive method, referred to as the four-point Wenner array electrical resistivity measurement (ERM), was developed based on the quantitative correlation between electrical resistivity and chloride diffusivity [2]. Since the electrical resistivity is a depth-related parameter, this technique has a potential application in damage detection [3].

Previous studies have, however, revealed that the ERM results may be unreliable because of the uncertain influence of many factors on measurements obtained. Consequently, extensive studies on the influence of individual factors have been reported to improve the reliability of ERM. Generally, the factors influencing ERM can be divided into two categories [4]: intrinsic factors (geometry of specimen [5,6], water to cement ratio [7], material types, aggregate size [8] and curing conditions [9,10]) and external factors (presence of steel rebar [11,12] and crack [13], probe spacing distance, testing temperature [14] and moisture content [15]).

Additional complexity arises due to the interrelationship of influencing factors. Limited studies have considered the combined influence of several factors. Liu [16] observed that the involvement of fly ash might result in higher electrical resistivity when considering the influence of curing temperature. Both specimen geometry and probe spacing were studied by Chen et al. [17] and it was found that their influence on the ERM greatly relied on the mutual geometrical correlation. Lubeck et al. [18] suggested that part of Portland cement could be replaced by slag to maintain similar mechanical and electrical properties with reduced cost. These previous studies have shown that it is not practical, due to time and cost implications, to employ experimental methods to correct the ERM results if the effect of all influential factors is to be considered simultaneously. To solve this problem, machine learning technology may be one of the potential solutions.

In the last decades, machine learning (ML) technology has been proved to be a powerful statistical tool to address non-linear regression and classification problems with multi-parameters [19]. ML technology can formulate potential correlations between input attributes and

* Corresponding author.
E-mail address: yimiao.huang@uwa.edu.au (Y. Huang).

https://doi.org/10.1016/j.autcon.2020.103155
Received 16 September 2019; Received in revised form 19 January 2020; Accepted 27 February 2020
Available online 06 March 2020

output targets by learning former occurrences without being explicitly programmed [20]. In civil engineering, many studies have been carried out to evaluate the performance of concrete structures by employing different types of ML algorithms [21,22]. Earlier research by Nehdi et al. [23] predicted the compressive strength of pre-formed foam cellular concrete using artificial neural networks (ANN). The same algorithm was applied by Fatih Altun et al. [24] to predict the compressive strength of steel fiber added lightweight concrete. Another highly integrated neural network model called extreme learning machine (ELM) was proposed by Yaseen et al. [25] to predict the compressive strength of lightweight foamed concrete. Prayogo et al. [26] applied a hybrid ML model composed of support vector machine (SVM) and symbiotic organisms search (SOS) models via their adaptive weightings for the prediction of the shear capacity of reinforced concrete. Bionic techniques like beetle antennae search (BAS) have been applied to tune the hyperparameters of the random forest model for predicting the uniaxial compressive strength of lightweight self-compacting concrete [27]. Recently, the potential of machine learning in assessing the durability and service-life of reinforced concrete structures has been analyzed by Taffese W et al. [22]. Meanwhile, they also studied the significant of parameters that control the chloride penetration in concrete through ensemble methods based on decision tree [28].

Apart from deep learning algorithms and other advanced hybrid models, enhanced boosting tree algorithms also have great potential in solving experimental problems [29]. For example, a boosting tree algorithm called XGBoost has shown superiorities in many data mining competitions in recent years due to its advantages [30] [31], which include (1) minimal requirements for attributes normalization, (2) processing missing values intelligently, (3) offering solutions to avoid overfitting [32]. Torlay et al. [33] tried to identify atypical language patterns to classify patients with epilepsy by XGBoost and obtained satisfactory results.

In the present study, an XGBoost-based prediction model is developed for normalizing the ERM results of concrete structures considering multi-factors simultaneously. 16 attributes that are classified into three categories (specimen parameters, curing conditions and testing parameters) are specified as input attributes for predicting the target after assessing all possible influential factors. A database containing 800 instances with complete data is created for training the XGBoost model based on previous experimental studies under different experimental circumstances. Statistical scores (RMSE, MAE and R^2) and the GridsearchCV function are used to evaluate and optimize the established model respectively. A ranking study of importance is also presented which identifies the critical attributes and their relative influence is further investigated by a variable control method through the proposed XGBoost model.

2. Methodology

2.1. Wenner array surface electrical resistivity measurement

Electrical resistivity discussed in this paper is measured using the four-point Wenner array meter. The principle of this type of ERM method is displayed in Fig. 1. Four equally spaced linear electrodes are aligned on the surface of the given specimen. Alternating current (AC) is applied passing through a certain depth of the specimen on two exterior electrodes while two interior electrodes measure the electric potential. The reason for using AC instead of direct current (DC) is that DC may result in inaccurate readings due to the polarization effect.

The corresponding electrical resistivity can be calculated by the following equation:

$$\rho = 2\pi a\frac{V}{I} \tag{1}$$

where ρ stands for the observed electrical resistivity; a denotes the distance between two close electrodes; V and I are the electric potential and current respectively. It can be noted that the Wenner array method is assumed to be performed on a semi-infinite geometry, which implies the probe spacing should be less than 1/4 of the specimen's height in order to ensure the detecting current passing through within the structure [34].

2.2. XGBoost algorithm

In this study, the XGBoost algorithm is selected for predicting concrete's electrical resistivity due to three main reasons. First, XGBoost is one of the most popular boosting tree algorithms for gradient boosting machine (GBM). It has been widely employed in industry due to its high performance in problem-solving and minimal requirement for feature engineering [35,36]. Second, compared with deep learning algorithms, XGBoost is recognized easier to use for small datasets running on CPU [37]. Considering the size of database in this study (800 instances), the XGBoost algorithm may be more appropriate than deep learning algorithms. Finally, the XGBoost was compared with Catboost and Keras neural network based on the database and results showed that the XGBoost had slightly better prediction accuracy than the other two.

Boosting tree algorithms are based on the decision tree, which is known as the classification and regression tree (CART). For regression tasks, CART divides the dataset into two subsets at each level according to the boundary for one variable until reaching the tree's maximum depth set by users. It can be described as below:

$$R_1(j,s) = \{x \mid x^j \le s\} \; and \; R_2(j,s) = \{x \mid x^j \ge s\} \tag{2}$$

Mean squared error of each leaf node is calculated:

$$MSE_{node} = \sum_{i \in node} (\hat{y}_{node} - y^{(i)})^2 \tag{3}$$

$$\hat{y}_{node} = \frac{1}{m_{node}} \sum\nolimits_{i \in node} y^{(i)} \tag{4}$$

where m_{node} is the number of instances in one node. The cost function for regression of CART can be expressed as:

$$\mathrm{J}(\mathrm{k}, t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right} \tag{5}$$

The algorithm will search for the best solutions of boundaries of variables to minimize the cost function. The prediction is then the average target value of all instances in one subset. However, CART trees prone to overfitting when dealing with regression tasks without regularization. One strategy for this issue is called ensemble by bagging a group of estimators, which, in this case, means multiple CART models.

The mechanism of XGBoost is keep adding and training new trees to fit residual errors of last iteration shown as Fig. 2. A predicted value is assigned to each instance by adding all corresponding leaves' scores together:

$$\hat{y} = \phi(x_i) = \sum_{k=i}^{K} f_k(x_i) \tag{6}$$

$$\text{and } f_k(x_i) = w_{q(x)}, \; f_k \in \mathscr{F} \tag{7}$$

where K denotes the quantity of trees; $f_k(x_i)$ stands for the outcome of input x_i for the kth tree; $w_{q(x)}$ specifies the score for each leaf node; $q(x)$ means the number of leaf nodes; $\mathscr{F}$ represents an assemble of all corresponding functions f_k.

The objective function of XGBoost contains two parts: the training error and the regularization, written as:

$$\mathrm{obj}(\theta) = \sum \mathrm{L}(\theta) + \sum \Omega(\theta) \tag{8}$$

where L is the loss function measuring the deviation of the predicted values from the actual values. Ω is the regularization function measuring the complexity of the training model in order to avoid overfitting.

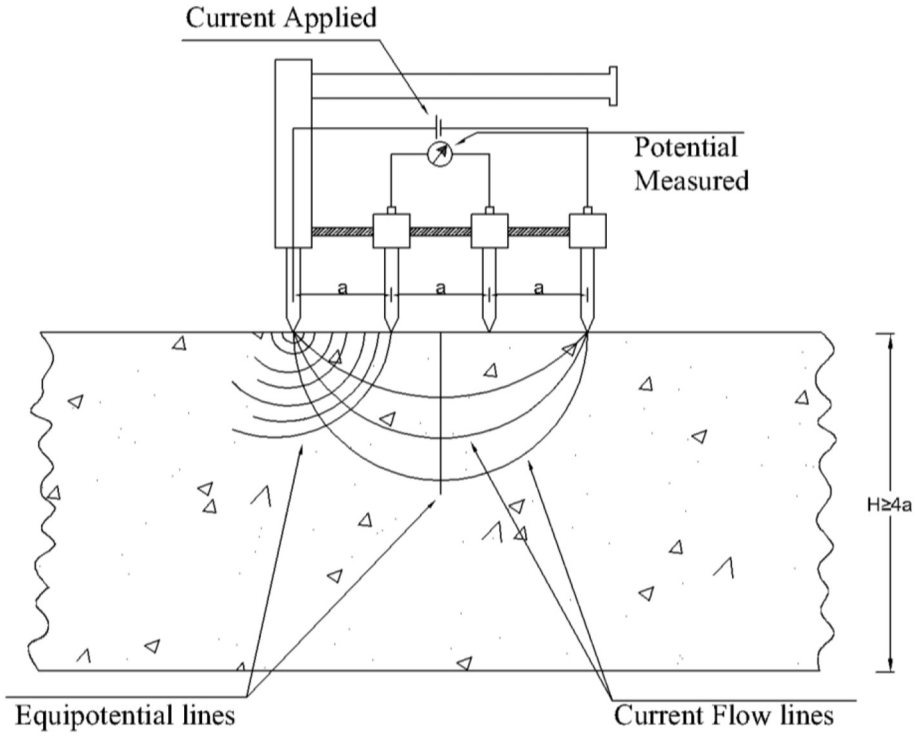


Fig. 1. Schematic for four-point Wenner array surface electrical resistivity measurement.

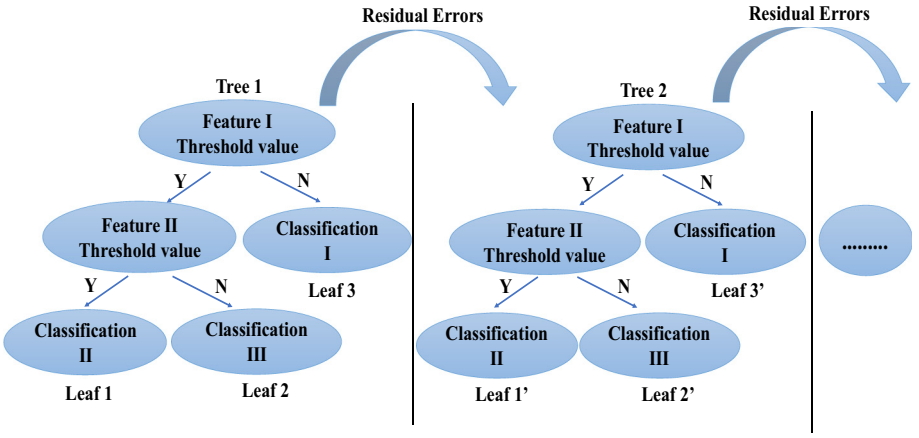


Fig. 2. Schematic of XGBoost Trees.

$$\Omega(\theta) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{9}$$

where T represents the total number of leaf nodes and ω is the score of each leaf node. γ and λ are controlling factors employed to avoid overfitting.

When a new tree is created to fit residual errors of last iteration, the predicted score for the tth tree can be expressed as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{10}$$

The objective function is thus rewritten as:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{11}$$

An appropriate function, f_t, is replaced with the second-order Taylor polynomial of $f_t = 0$. Accordingly, the objective function can be approximated as:

$$\mathscr{L}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{12}$$

where g_i is the first-order derivative and h_i denotes the second-order derivative:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{13}$$

Since previous (t − 1) trees' residual errors (y) have minimal influence on the modification of the objective function, Eq. (11) is then simplified as:

$$\widetilde{\mathscr{L}^{(t)}} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{14}$$

As each instance will finally be classified into one leaf node, all instances that belong to the same leaf node can be reassembled as:

$$\text{obj}^{(t)} \approx \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma \tag{15}$$

Therefore, the optimum w and objective function are derived as:

Table 1
Results of attributes filtering.

Class	Category	Original attributes	Decided attributes	
Input attributes	Specimen parameters	Measured sectional area (mm∗mm)	Length (mm)	
			Height (mm)	
		w/c ratio	w/c ratio	
		Cement content (kg/m^3)	Cement content (kg/m^3)	
		Fine/coarse aggregate content (kg/m^3)	×	
		Extra fillers content (kg/m^3)	Silica fume content (kg/m^3)	
			Fly ash content (kg/m^3)	
			Slag content (kg/m^3)	
		Maximum aggregate size (mm)	Maximum aggregate size (mm)	
	Curing conditions	Curing environment	Stage one	Curing environment1
				Curing temperature1 (°C)
		Curing temperature (°C)		Curing age (day)
			Stage two	Curing environment2
		Curing age (day)		Curing temperature2 (°C)
				Start date (day)
				End date (day)
	Testing parameters	Testing temperature (°C)	×	
		Specimen saturation	×	
		Electrode spacing (mm)	Electrode spacing (mm)	
		Presence of steel rebar/crack	×	
Output target		Electrical resistivity (kΩ·cm)	Electrical resistivity (kΩ·cm)	

Table 2
Resources of data collection.

References	Specimen parameters							Curing conditions	Electrode spacing	Electrical resistivity	Data collected
	Length and height	Materials									
		w/c ratio	Cement content	Fly ash content	Slag content	Silica fume content	Maximum aggregate size				
Sengul [6]	√	√	√	√	×	×	√	√	√	√	39
Liu [16,42]	√	√	√	√	√	×	√	√	√	√	72
Chen [17]	√	√	√	×	×	×	√	√	√	√	24
Lübeck [18]	√	√	√	×	√	×	√	√	√	√	30
Rupnow [38]	√	√	√	√	√	√	√	√	√	√	64
Su [43]	√	√	√	×	×	×	√	√	√	√	3
Sanchez Marquez [44]	√	√	√	√	√	×	√	√	√	√	136
Shahroodi [45]	√	√	√	√	×	√	√	√	√	√	396
Güneyisi [46]	√	√	√	×	×	×	√	√	√	√	12
Ramezanianpour [47]	√	√	√	×	×	×	√	√	√	√	24

√: data available; ×: data missing.

$$w_j^* = -\frac{G_j}{\mathrm{H}_j + \lambda} \quad obj = -\frac{1}{2}\sum\nolimits_{j=1}^{t} \frac{G_j^2}{H_j + \lambda} + \gamma \mathrm{T} \tag{16}$$

XGBoost's superiority lies in its reliable objective function for creating trees. Meanwhile, it serves several effective parameters such as maximum depth and splitting threshold. XGBoost also applies two overfitting-avoid strategies: shrinkage and column subsampling. Shrinkage specifies that all leaves' scores will multiply a shrinkage weight η in every iteration to ensure that one tree's influence on the model won't be too large and allow more space for the following trees. Column subsampling enables to build a tree with only a part of attributes in the same way as random forest.

3. XGBoost model establishment

3.1. Database setting up

The purpose of this paper is to describe the establishment of an XGBoost prediction model to solve the ERM normalization problem for concrete structures considering multi-factors. It is necessary to set up an original and experimental database containing instances with all potential input attributes related to the predicting target - electrical resistivity.

All these attributes are analyzed before the database being employed to train the XGBoost model to filter some attributes that have little influence on the ERM results. Some others that are too abstract to be recognized by machine learning algorithms are also processed. The stepwise analysis is elaborated as below:

1. For some elements, the most influential feature can only be selected through previous studies. For example, aggregate has quite a few features such as type, distribution, size and content. Some of the features may be too complicated to organize in the database and the results may be invalid if all the features are considered [38]. Therefore, the maximum size is chosen as aggregate's representative feature in this study.
2. For the testing temperature, a large quantity of studies has proved that the correlation between electrical resistivity and testing temperature complies with the Archie and Arrhenius law. An accurate equation has been proposed for resistivity versus temperature normalization based on the activation energy [39]. Therefore, it is not necessary to concern it in this study.
3. Specimen saturation is another attribute that can be left out because it is determined by curing conditions. Otherwise its contributions to ERM results will be overlapped by that of curing condition data.
4. The influence of reinforcement and cracks is not considered in this

Table 3
Data of some collected instances.

Length	Height	w/c ratio	Cement content	Fly ash content	Slag content	Silica fume content	Maximum aggregate size	Curing environment1	Curing temperature1	Curing age	Curing environment2	Curing temperature2	Start date	End date	Electrode spacing	Electrical resistivity
406	75	0.41	354.88	0	0	0	10	Moisture	23	7	Moisture	23	0	0	50	5
406	75	0.56	265.85	0	89.02	0	10	Moisture	23	14	Moisture	23	0	0	40	15.7
170	100	0.3	517	0	0	0	19	Moisture	23	28	Moisture	23	0	0	30	25.18
406	75	0.48	282.73	0	102.44	24.59	10	Moisture	23	7	Moisture	23	0	0	20	17
170	100	0.3	517	0	0	0	19	Moisture	23	91	Moisture	23	0	0	30	35.29
406	75	0.46	281.71	0	93.9	0	10	Moisture	23	91	Moisture	23	0	0	40	76.4
200	100	0.42	276	0	100	24	10	Moisture	23	56	Moisture	23	0	0	50	211.9
200	100	0.45	354.88	0	0	0	10	Moisture	23	28	Moisture	23	0	0	25	7.8
200	100	0.4	346.86	76.14	0	0	16	Water	20	7	Water	5	8	16	20	10.26
200	100	0.56	265.85	0	89.02	0	10	Moisture	23	56	Moisture	23	0	0	30	29.3
170	100	0.55	306	0	0	0	19	Moisture	23	28	Moisture	23	0	0	30	9.76
200	100	0.52	340	0	0	0	14	Lime water	23	70	Moisture	23	8	28	38	8.19
406	75	0.55	265.85	0	89.02	0	10	Moisture	23	7	Moisture	23	0	0	30	6
170	100	0.42	415	0	0	0	19	Moisture	23	91	Moisture	23	0	0	30	14.31
300	120	0.4	170	100	70	0	14	Water	23	134	Moisture	23	8	28	38	106.23
200	100	0.46	281.71	0	93.9	0	10	Moisture	23	28	Moisture	23	0	0	30	37.4
200	100	0.45	170	70	100	0	14	Lime water	23	10	Moisture	23	0	0	38	3.01

paper as these two variables are too complicated to control in experimental tests, leading to a lack of data. It is also difficult to formulate their character in uniform data forms in the program.

5. The influence of geometry on the ERM is determined by the geometric relationship between the probe spacing and specimen's dimensions. Therefore, the geometry attribute can be sorted into two attributes: length and height in order to make machine learning algorithms discern this type of attribute.
6. Attributes pertaining to curing conditions are all divided into two categories as most experiments differentiate the standard curing stage from other curing stage. Since the period of the second curing stage is determined by that of the first one, it is represented by the start and end date to maintain date consistency.

Details of attributes that may influence the ERM are summarized in Table 1.

The next step is to collect experimental data from previous studies. In this research, a database containing 800 instances was assembled primarily based on available research articles shown in Table 2. Although individual research could only account for a limited number of attributes while other attributes fixed with certain values, a wide range of data for each attribute are available by cross-correlating the limited attributes. A subset of the dataset is displayed in Table 3 in dataframe structure.

It is seen in the database that electrical resistivity values vary from 2 to maximum of 320 kΩ·cm under the comprehensive influence of multi-attributes. The specific influence of the individual attribute on the measurement results cannot be differentiated from the original data.

3.2. Data preprocessing

After importing the database into the program, the first step involves shuffling the database to a random order before splitting it into the training and testing subsets based on an acknowledged ratio of 0.8/0.2 to avoid the human interference referred to as data snooping bias. The training set is used for establishing the XGBoost model, while the testing set evaluates the prediction accuracy of the established model. K-fold cross validation method is employed to optimize the training process by dividing the training set into multiple subsets.

Before training an XGBoost model, data preprocessing works (or the feature engineering) are necessary to improve the training performance. This step is significant for the ML algorithm to 'understand' the data better. In fact, some attribute processing works have already been carried out during the data collection process. Therefore, there are only a limited number of attributes that need to be interpreted, especially those attributes belonging to the curing conditions because data types of other attributes are already numeric (see Table 3).

Since the predictive target is the electrical resistivity, the 'Electrical resistivity' column should be put aside as the 'label' of training instances.

It should be noted that values of the curing environment attribute are categorical which are 'moisture', 'lime water', 'water' and 'air'. In order to interpret these data into numeric forms, factorization method is applied. Data belonging to two curing environment attributes are then combined as one curing environment strategy. Once being factorized, the curing environment data become numerals. However, there is practically no quantitative relationship between different categories. In case that the XGBoost algorithm would be confused by this false relationship, a common solution is to interpret those factorized data into one binary attribute per category further by the One-hot-encoding function. In this case, for example, strategy [moisture, moisture] is interpreted into [1, 0, 0, 1, 0, 0, 0].

Although the curing temperature data are numeric, they are not in sequential order as most researchers prefer to set curing temperature at the room temperature (23 ± 2 °C) following standards such as ASTM C39 [40] and ASTM C1202 [41]. Researchers studying how curing

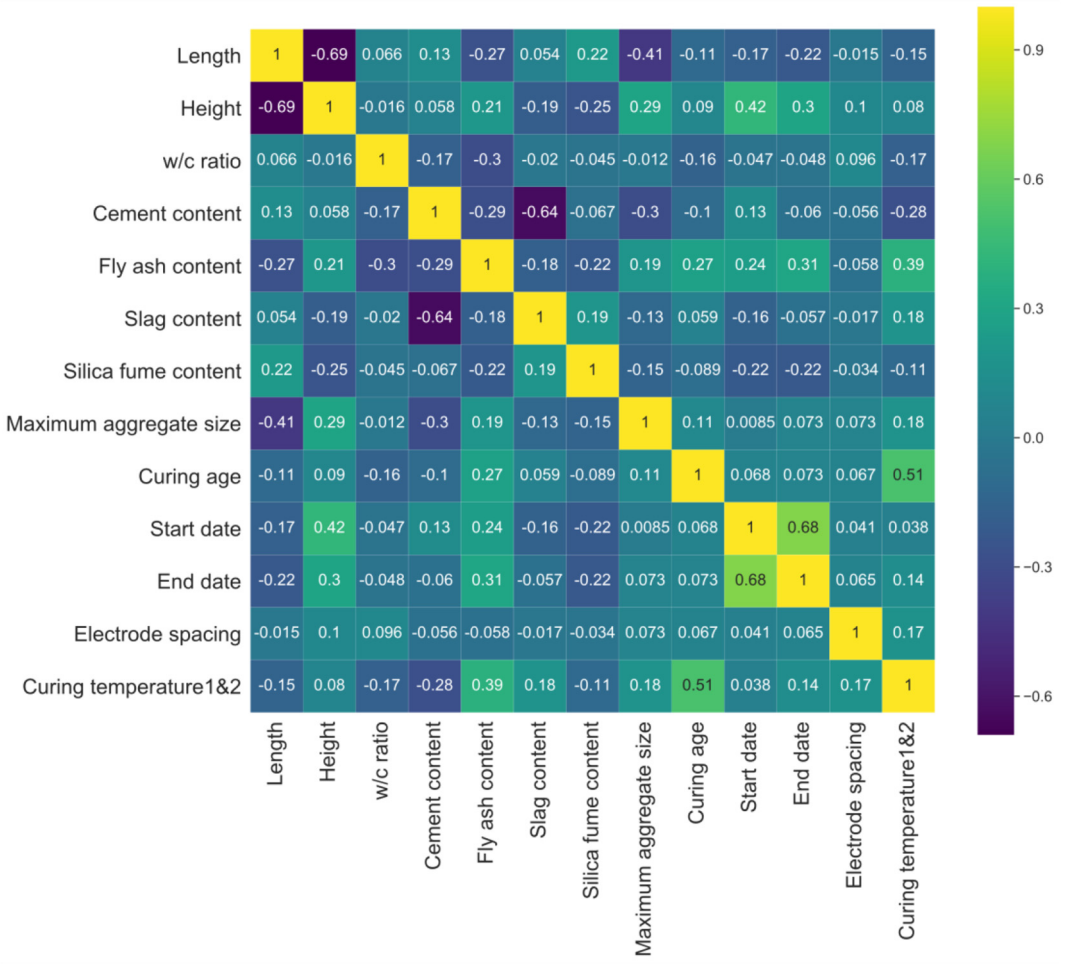


Fig. 3. Pearson correlations of each two attributes.

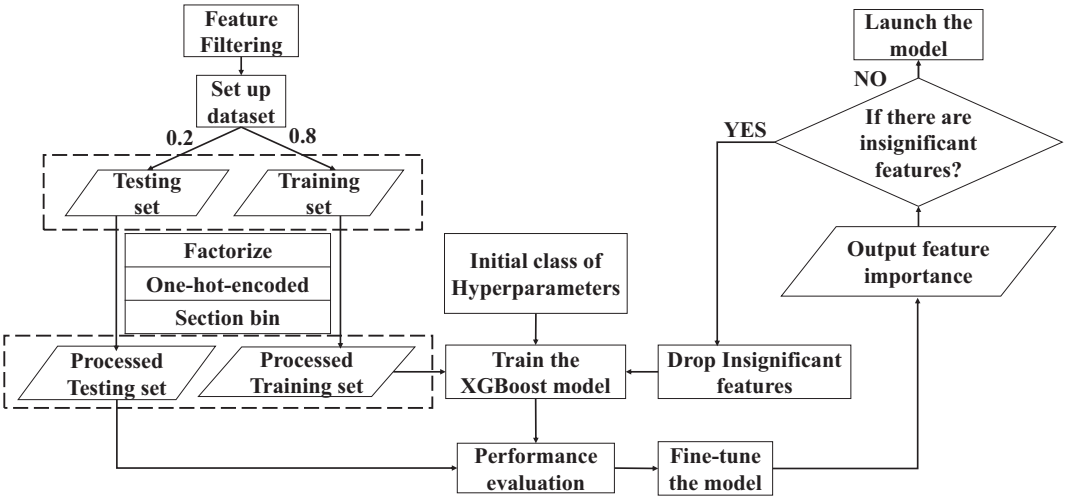


Fig. 4. Flow chart for establishment of the XGBoost model.

Table 4
Initial values of hyperparameters for XGBoost model.

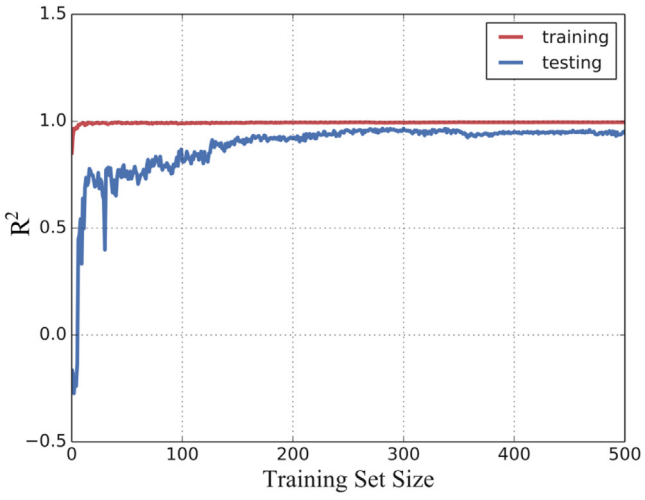
max_depth	learning_rate	n_estimatiors	gamma	subsample
10	0.3	100	0.05	0.6

temperature affect the ERM would like to control the temperature by putting specimens in a freezer chamber or a hot oven. To fix this data-scattering issue, the solution is to classify the curing temperature data into three categorical bins: low temperature, room temperature and high temperature, followed by similar measures implemented on the curing environment data.

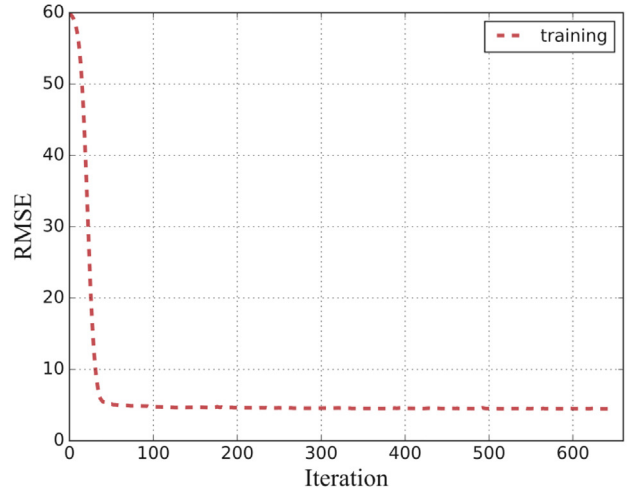
Following the foregoing steps, all necessary data are preprocessed to

Table 5
Ranges of hyperparameters for model fine-tuning.

Name	Range of values	Common difference
max_depth	1–15	1
learning_rate	0.1–0.5	0.05
n_estimators	200–1000	50
gamma	0.01–0.05	0.01
subsample	0.1–1.5	0.1



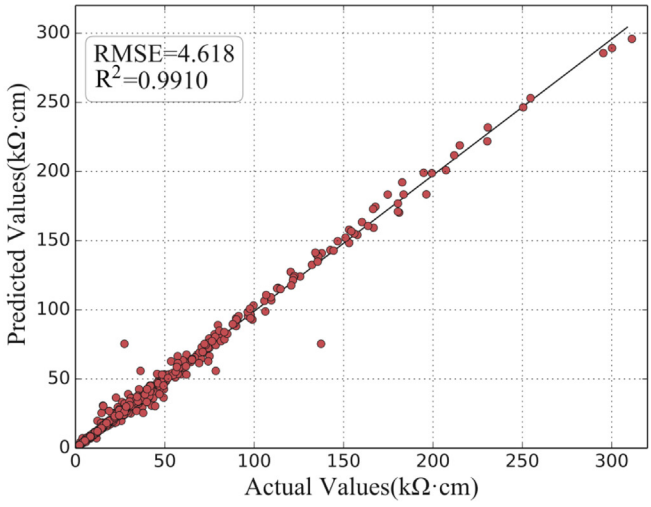
(a) R^2 versus training set size



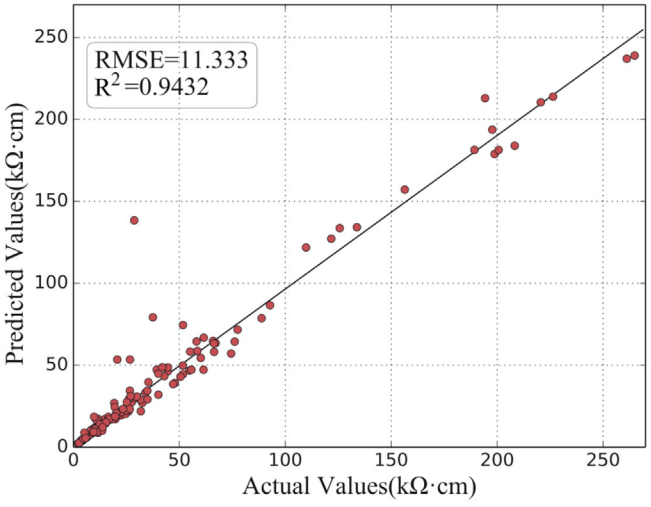
(b) RMSE versus iteration

Fig. 5. Learning curves for the training process.

be recognizable for the algorithm and an effective database is ready for training the model after replacing the old data with the processed data. A preliminary understanding of the database can be achieved by calculating the Pearson correlation coefficient (PCC) between each two attributes as shown in Fig. 3. Pearson correlation coefficient measures the level of linear correlations between two variables. It ranges from −1 to 1 and coefficient close to zero means that there is no linear correlation. If two attributes prove linear correlated with each other according to the PCC, it means one of them can be neglected to optimize the dataset. Since all calculated PCCs of attributes from the database are quite small (most less than 0.5), the influence of determined attributes on the ERM can be assumed to be independent of each other.



(a) training set



(b) testing set

Fig. 6. Correlations between predicted values versus actual values.

3.3. Model training

The program is developed in a Python virtual environment. The procedure of the XGBoost model establishment is explained in Fig. 4.

According to the XGBoost's principle, some hyperparameters are essential towards the training performance of the XGBoost algorithm including max_depth, learning_rate, n_estimators, gamma, subsample and objective, among which objective is used to specify the learning task for choosing the right objective function. In this study, 'regression: gamma' is determined for hyperparameter objective because it is a typical regression task. 'gamma' is the minimum loss reduction required to make a further partition on a leaf node of the tree.

Random initial values for other hyperparameters are listed in Table 4 according to former experience. max_depth means the maximum tree depth. learning_rate defines the boosting learning rate. n_estimators determines the number of trees to fit. Subsample is the subsample ratio of all instance for partial training.

Once setting the initial values for hyperparameters to train the XGBoost algorithm, a corresponding model is generated based on the preprocessed training set. During this process, K-fold cross-validation

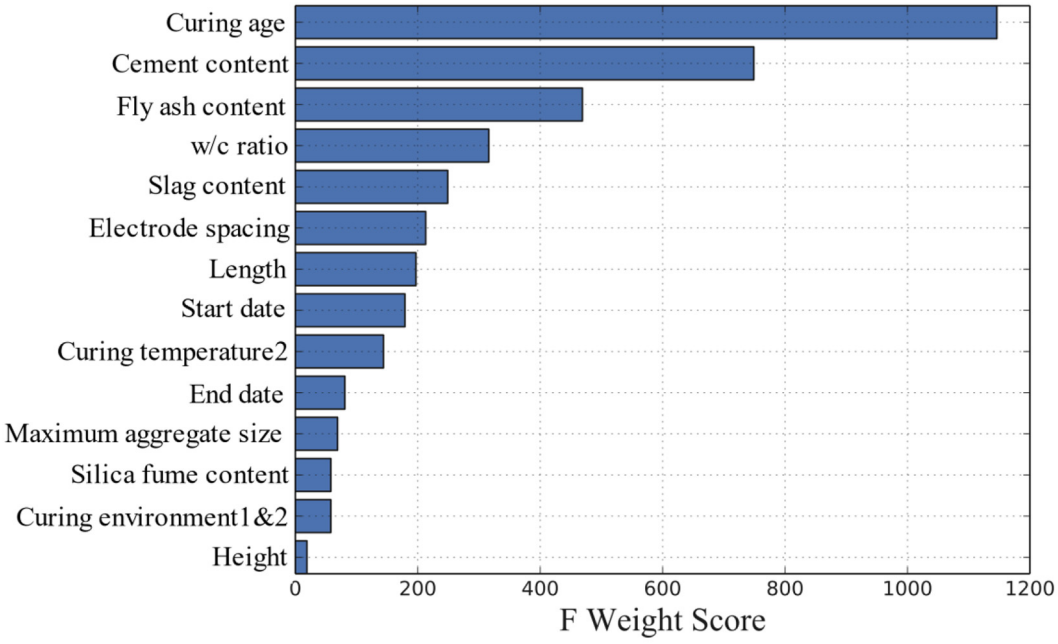


Fig. 7. Importance rank of attributes.

method is employed to improve the training performance by randomly splitting the training set into k distinct subsets called folds. It then trains and evaluates the established XGBoost model k times, picking one of folds for evaluation every time and training on the other (k − 1) folds. Finally, an array containing k evaluation scores is obtained, based on which the model would be further optimized.

After the model is constructed, comparison between the predicted and observed electrical resistivity values is carried out in conjunction with statistical metrics including RMSE, MAE and R^2 (as defined below) to appraise the accuracy of the established XGBoost model. These three scores are basic evaluation indexes for regression tasks.

Root mean square error (RMSE) represents the errors between predicted values and actual values and is calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^* - y_i)^2} \tag{17}$$

Mean absolute error (MAE) is the average of absolute errors of predicted values, determined from:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i^* - y_i| \tag{18}$$

R-Square (R^2) measures the deviation of a group of data and represents the goodness of regression fitting:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i^* - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2} \tag{19}$$

where N stands for the number of all instances; y_i^* and y_i are the respective predicted and observed resistivity values; $\bar{y}$ denotes the mean of predicted resistivity values.

The established initial XGBoost model is then fine-tuned by searching for the best combination of hyperparameters. The hyperparameters that are vital to the establishment of model should be identified and a suitable range of values should be examined as shown in Table 5. In order to narrow down the search range and improve the fine-tuning efficiency, rough search is conducted on a large range of hyperparameters with loose common difference based on initial values at first. Then a fuzzy range for each hyperparameter is identified. Common difference guarantees the searching precision. The program then evaluates all the possible combinations of hyperparameter values.

The best estimator as well as its hyperparameters are finally achieved. The identified optimized model is then applied to the testing set to assess its predictive performance. Following that, the XGBoost program analyzes the importance weight of each attribute. If some attribute is assumed to be less important towards the prediction, it will be dropped for the next training iteration.

A deep understanding of significant attributes affecting the electrical resistivity measurement results can be obtained by deploying the proposed XGBoost model upon a set of artificial datasets. In the artificial dataset, the studied attribute should involve a wide range of values while all the other attributes' values fixed.

4. Results and discussion

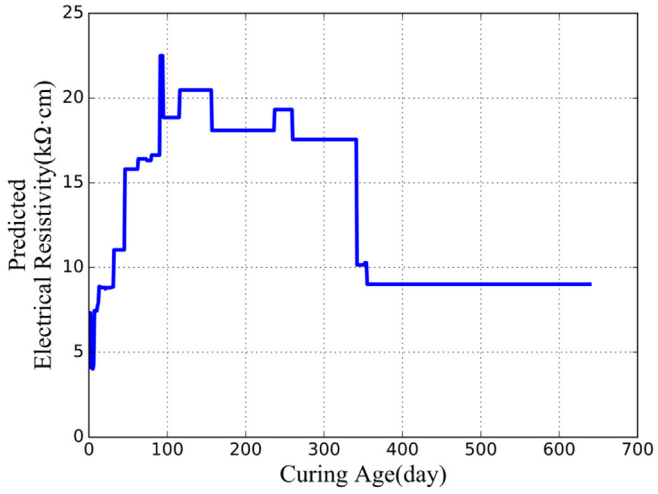
4.1. Training results of the proposed model

Once the best combination of hyperparameters is identified, the best estimator model is achieved. It is then evaluated with the training set and the testing set respectively. The two running processes are illustrated in Fig. 5. Fig. 5(a) shows the development of the R^2 score for both training and testing with the increasing size of training set. The training curve reaches its largest R^2 score of almost 1 after a few running loops and it remains stable for the rest of running procedure. The testing curve increases steeply to 0.63 during the initial 50 loops. After that, the growing slope decreases steadily to stabilize at a score of around 0.93.

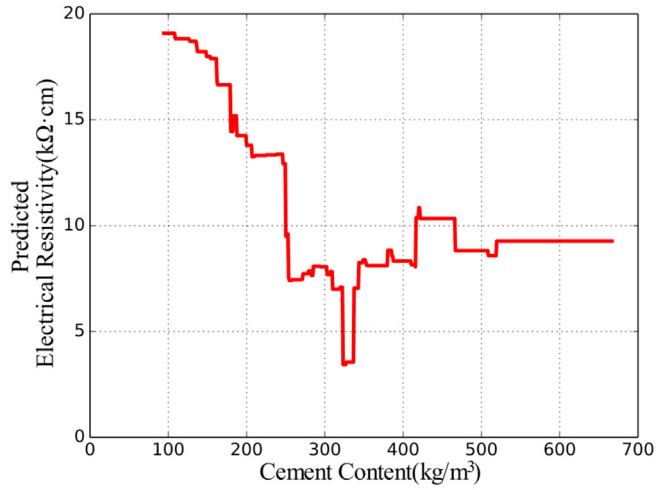
According to the learning curve in Fig. 5(b), the proposed XGBoost model converges to the minimum RMSE score quickly within the first 50 iterations and then maintains constantly. Based on the above analyses, the proposed model proves to be viable and reliable not only because of the satisfying evaluation scores in both training and testing, but also because of a very small performance gap between the training and testing indicating no overfitting issue.

4.2. Predictive performance of proposed model

Predicted electrical resistivity values are derived by deploying the proposed model on the training set and the testing set respectively. Scatter points of predicted values versus actual values as well as their



(a) Curing age



(b) Cement content

Fig. 8. Specific influence of two dominating factors on ERM.

linear regression fitting lines are plotted in Fig. 6. Statistical scores including RMSE and R^2 are calculated for evaluating the goodness of the fit. Very satisfactory coefficients of both regression lines can be observed (0.991 and 0.943) indicating a high fitting performance. However, even if the proposed model shows a high predictive accuracy as credited by large R^2 and low RMSE, the predictive performance on the testing set is slightly worse than that on the training set. This is probably because the value range of the testing set is mainly focused on the section under 100 kΩ·cm leading to a serious imbalance ratio on the data range due to its small dataset size.

4.3. Importance of attributes

The importance rank of all attributes is deduced by the program based on their F weight scores contributing to the establishment of the proposed model. The F weight score of one attribute is calculated based on the total times of it being chosen for producing leaf nodes during XGBoost trees training. Fig. 7 shows that the two most contributing attributes are curing age and cement content. The addition of fly ash is the third important attribute affecting the ERM followed by w/c ratio.

The two most important attributes are analyzed further to study the patterns of their influence on the electrical resistivity results respectively. A dataset containing synthetic data with a large range of sequential values for the studied attribute while fixed values for the rest of attributes is created and imported into the proposed XGBoost model. Fig. 8 displays the predicted results of each attribute respectively. From Fig. 8(a), except for a slight drop at the beginning, electrical resistivity grows steeply from 4.5 to 16 kΩ·cm during the early developing stage of 28 days. It then increases marginally by about 1.5 kΩ·cm by around 90 days. After about 90 days, there is another rapid growth in electrical resistivity, which lasts for almost 20 days before there is a slight drop to 18.6 kΩ·cm, remaining constant at this value until 360 days. The reason behind this phenomenon may be that the development of hydration in concrete is beneficial for the electrical resistivity growth due to that denser concrete leads to lower porosity. In addition, there are two rapid growing stages during the process. Once the concrete is fully cured under the similar environment, the growth of its electrical resistivity will almost stop.

The electrical resistivity and cement content can be correlated quantitatively. By the amount of around 200 kg/m^3, electrical resistivity decreases gradually to 13 kΩ·cm with increasing cement content and keeps constant for the next 50 kg/m^3 growing phase. Since then, however, it experiences a sharp drop to almost 2 kg/m^3 with fluctuations and climbs back to the 8–9 kg/m^3 level for the rest of cement content values. It can be explained by that the increasing content of cement is beneficial for the electrical conductivity development. More intense hydration may be caused with more cement content leading to high temperature difference between the inside and the outside of the specimen due to the heat released. During this process, more microstructures can be formed, which enables electron particles to pass through the concrete leading to a lower electrical resistivity.

In summary, longer curing age is beneficial for the development of concrete's electrical resistivity while an opposite trend happens with more cement content. Both factors meet a threshold, over which their influence on the measured results can be very slight. On the other hand, with certain content of cement, the first hundred days of curing period is significant for ensuring enough electrical resistivity to contain the chloride diffusivity and should be cared more about.

5. Conclusions

Electrical resistivity measurement (ERM) is susceptible to significant uncertainties due to the influence of many factors. The present research is an early study for using a machine learning technique called XGBoost algorithm to establish a predictive model in order to normalize Wenner array ERM results of concrete structures by given parameters.

According to the evaluation scores of high coefficients of regression fitting lines (0.991 and 0.943) and low RMSE values (4.6 and 11.3 kΩ·cm) for both training and testing datasets, conclusion can be drawn that the proposed XGBoost model achieves satisfactory predictive capability for Wenner array ERM of concrete structures considering multi-factors simultaneously. The XGBoost model therefore offers a reliable and intelligent approach for normalizing the observed ERM results to values at a reference condition. It can also be utilized to predict and assess the durability of concrete structures. Based on the rank of attributes' importance scores, curing age and cement content are the two most contributing factors (almost 2.8 and 1.8 times of that of the third important factor respectively). Their relative influence on ERM is investigated through the variable control analysis, which is beneficial for future experimental studies. Both attributes have a threshold, over which their influence on ERM results tends to be very slight.

Additionally, the proposed model in this paper works comparatively unsatisfactory in the section of high electrical resistivity due to lack of relevant data. The predictive ability of the proposed model can be improved with an extended dataset that covers a wider range of experimental data. The influence of reinforcement and cracks on ERM will be taken into consideration in the next stage of this research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Experimental data used in machine learning modeling in this paper are retrieved from open literatures. All data sources are acknowledged.

This research is founded by the Discovery Projects of the Australian Research Council (Grant No. DP180104035).

The author wants to acknowledge the support of University of Western Australia through a 'Scholarship for International Research Fees and Ad Hoc Postgraduate Scholarship'.

References

[1] J.P. Lynch, C.R. Farrar, J.E. Michaels, Structural health monitoring: technological advances to practical implementations [scanning the issue], Proceedings of the Institute of Electrical and Electronics Engineers 104 (8) (2016) 1508–1512, https://doi.org/10.1109/JPROC.2016.2588818.

[2] O. Sengul, Use of electrical resistivity as an indicator for durability, Constr. Build. Mater. 73 (2014) 434–441, https://doi.org/10.1016/j.conbuildmat.2014.09.077.

[3] N. Wiwattanachang, P.H. Giao, Monitoring crack development in fiber concrete beam by using electrical resistivity imaging, J. Appl. Geophys. 75 (2) (2011) 294–304, https://doi.org/10.1016/j.jappgeo.2011.06.009.

[4] P. Azarsa, R. Gupta, Electrical resistivity of concrete for durability evaluation: a review, Adv. Mater. Sci. Eng. (2017), https://doi.org/10.1155/2017/8453095.

[5] R. Spragg, Y. Bu, K. Snyder, D. Bentz, J. Weiss, Electrical Testing of Cement-Based Materials: Role of Testing Techniques, Sample Conditioning, and Accelerated Curing, (2013), https://doi.org/10.5703/1288284315230.

[6] O. Sengul, O.E. Gjørv, Electrical resistivity measurements for quality control during concrete construction, American Concrete Institute Materials Journal 105 (6) (2008) 541, https://doi.org/10.14359/20195.

[7] T.D. Rupnow, P. Icenogle, Evaluation of Surface Resistivity Measurements as an Alternative to the Rapid Chloride Permeability Test for Quality Assurance and Acceptance (No. FHWA/LA. 11/479), Louisiana Transportation Research Center, 2011 Website, (Accessed date: 16/1/2020) https://rosap.ntl.bts.gov/view/dot/22099.

[8] W. Morris, E.I. Moreno, A.A. Sagüés, Practical evaluation of resistivity of concrete in test cylinders using a Wenner array probe, Cem. Concr. Res. 26 (12) (1996) 1779–1787, https://doi.org/10.1016/S0008-8846(96)00175-5.

[9] F. Presuel-Moreno, Y.Y. Wu, Y. Liu, Effect of curing regime on concrete resistivity and aging factor over time, Constr. Build. Mater. 48 (2013) 874–882, https://doi.org/10.1016/j.conbuildmat.2013.07.094.

[10] J. Weiss, K. Snyder, J. Bullard, D. Bentz, Using a saturation function to interpret the electrical properties of partially saturated concrete, J. Mater. Civ. Eng. 25 (8) (2012) 1097–1106, https://doi.org/10.1061/(ASCE)MT.1943-5533.0000549.

[11] O. Sengul, O.E. Gjørv, Effect of embedded steel on electrical resistivity measurements on concrete structures, American Concrete Institute Materials Journal 106 (1) (2009) 11, https://doi.org/10.14359/56311.

[12] J. Sanchez, C. Andrade, J. Torres, N. Rebolledo, J. Fullea, Determination of reinforced concrete durability with on-site resistivity measurements, Mater. Struct. 50 (1) (2017) 41, https://doi.org/10.1617/s11527-016-0884-7.

[13] A.A. Shah, Y. Ribakov, Non-destructive measurements of crack assessment and defect detection in concrete structures, Mater. Des. 29 (1) (2008) 61–69, https://doi.org/10.1016/j.matdes.2006.12.002.

[14] Y. Liu, F.J. Presuel-Moreno, Normalization of temperature effect on concrete resistivity by method using Arrhenius law, American Concrete Institute Materials Journal 111 (4) (2014) 433–442, https://doi.org/10.14359/51686725.

[15] C.K. Larsen, E.J. Sellevold, F. Askeland, J.M. Østvik, O. Vennesland, Electrical resistivity of concrete part II: influence of moisture content and temperature, 2nd International Symposium on Advances in Concrete through Science and Engineering, Quebec, Canada, 2006 Website https://pdfs.semanticscholar.org/246b/8f5e2750abcb411e1871a713362635613a5e.pdf (Accessed date: 16/1/2020).

[16] Y. Liu, F. Presuel-Moreno, Effect of elevated temperature curing on compressive strength and electrical resistivity of concrete with fly ash and ground-granulated blast-furnace slag, American Concrete Institute Materials Journal 111 (5) (2014), https://doi.org/10.14359/51686913.

[17] C.T. Chen, J.J. Chang, W.C. Yeih, The effects of specimen parameters on the resistivity of concrete, Constr. Build. Mater. 71 (2014) 35–43, https://doi.org/10.1016/j.conbuildmat.2014.08.009.

[18] A. Lübeck, A.L.G. Gastaldini, D.S. Barin, H.C. Siqueira, Compressive strength and electrical properties of concrete with white Portland cement and blast-furnace slag, Cem. Concr. Compos. 34 (3) (2012) 392–399, https://doi.org/10.1016/j.cemconcomp.2011.11.017.

[19] D. Prayogo, M.Y. Cheng, Y.W. Wu, D.H. Tran, Combining machine learning models via adaptive ensemble weighting for prediction of shear capacity of reinforced-concrete deep beams, Eng. Comput. (2019) 1–19, https://doi.org/10.1007/s00366-019-00753-w.

[20] A. Géron, Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, Inc, 2017 (ISBN: 1491962267, 9781491962268).

[21] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Application of machine learning to construction injury prediction, Autom. Constr. 69 (2016) 102–114, https://doi.org/10.1016/j.autcon.2016.05.016.

[22] W.Z. Taffese, E. Sistonen, Machine learning for durability and service-life assessment of reinforced concrete structures: recent advances and future directions, Autom. Constr. 77 (2017) 1–14, https://doi.org/10.1016/j.autcon.2017.01.016.

[23] M. Nehdi, Y. Djebbar, A. Khan, Neural network model for preformed-foam cellular concrete, Dent. Mater. J. 98 (5) (2001) 402–409, https://doi.org/10.14359/10730.

[24] F. Altun, Ö. Kişi, K. Aydin, Predicting the compressive strength of steel fiber added lightweight concrete using neural network, Comput. Mater. Sci. 42 (2) (2008) 259–265, https://doi.org/10.1016/j.commatsci.2007.07.011.

[25] Z.M. Yaseen, R.C. Deo, A. Hilal, A.M. Abd, L.C. Bueno, S. Salcedo-Sanz, M.L. Nehdi, Predicting compressive strength of lightweight foamed concrete using extreme learning machine model, Adv. Eng. Softw. 115 (2018) 112–125, https://doi.org/10.1016/j.advengsoft.2017.09.004.

[26] D. Prayogo, M.Y. Cheng, Y.W. Wu, D.H. Tran, Combining machine learning models via adaptive ensemble weighting for prediction of shear capacity of reinforced-concrete deep beams, Eng. Comput. (2019) 1–19, https://doi.org/10.1007/s00366-019-00753-w.

[27] J. Zhang, G. Ma, Y. Huang, F. Aslani, B. Nener, Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression, Constr. Build. Mater. 210 (2019) 713–719, https://doi.org/10.1016/j.conbuildmat.2019.03.189.

[28] W.Z. Taffese, E. Sistonen, Significance of chloride penetration controlling parameters in concrete: ensemble methods, Constr. Build. Mater. 139 (2017) 9–23, https://doi.org/10.1016/j.conbuildmat.2017.02.014.

[29] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016 (ISBN: 0128043571, 9780128043578).

[30] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 2016, pp. 785–794, , https://doi.org/10.1145/2939672.2939785 August.

[31] Ron Bekkerman, The present and the future of the Knowledge, Discovery and Data Mining Cup Competition: an outsider's perspective, 2015. Website: https://www.linkedin.com/pulse/present-future-kdd-cup-competition-outsiders-ron-bekkerman/ (Accessed date:16/1/2020).

[32] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232 Website https://www.jstor.org/stable/2699986 (Accessed date: 16/1/2020).

[33] L. Torlay, M. Perrone-Bertolotti, E. Thomas, M. Baciu, Machine learning–XGBoost analysis of language networks to classify patients with epilepsy, Brain Informatics 4 (3) (2017) 159, https://doi.org/10.1007/s40708-017-0065-7.

[34] R.B. Polder, Test methods for on site measurement of resistivity of concrete—a RILEM TC-154 technical recommendation, Constr. Build. Mater. 15 (2–3) (2001) 125–131, https://doi.org/10.1016/S0950-0618(00)00061-1.

[35] A. Möller, V. Ruhlmann-Kleider, C. Leloup, J. Neveu, N. Palanque-Delabrouille, J. Rich, R. Carlberg, C. Lidman, C. Pritchet, Photometric classification of type Ia supernovae in the SuperNova Legacy Survey with supervised learning, J. Cosmol. Astropart. Phys. (12) (2016) 8, https://doi.org/10.1088/1475-7516/2016/12/008.

[36] D. Tamayo, A. Silburt, D. Valencia, K. Menou, M. Ali-Dib, C. Petrovich, C.X. Huang, H. Rein, C. van Laerhoven, A. Paradise, A. Obertas, A machine learns to predict the stability of tightly packed planetary systems, The Astrophysical Journal Letters 832 (2) (2016) L22, https://doi.org/10.3847/2041-8205/832/2/L22.

[37] Gad Benram, XGBoost or TensorFlow? Website https://blog.doit-intl.com/xgboost-or-tensorflow-63f4c92d4377, (2018) (Accessed date: 16/1/2020).

[38] T.D. Rupnow, P. Icenogle, Evaluation of surface resistivity measurements as an alternative to the rapid chloride permeability test for quality assurance and acceptance (no. FHWA/LA. 11/479), Website Louisiana Transportation Research Center, https://rosap.ntl.bts.gov/view/dot/22099, (2011).

[39] T.M. Chrisp, G. Starrs, W.J. McCarter, E. Rouchotas, J. Blewett, Temperature-conductivity relationships for concrete: an activation energy approach, J. Mater. Sci. Lett. 20 (12) (2001) 1085–1087, https://doi.org/10.1023/A:1010926426753.

[40] ASTM (American Society for Testing and Materials), C, Standard test method for compressive strength of cylindrical concrete specimens, ASTM C39/C39M-12. Website https://www.astm.org/DATABASE.CART/HISTORICAL/C39C39M-12.htm, (2012) (Accessed date: 16/1/2020).

[41] . ASTM (American Society for Testing and Materials), C., Standard test method for electrical indication of concrete's ability to resist chloride ion penetration, Annual Book of ASTM Standards, 2012. Website: https://www.astm.org/DATABASE.CART/HISTORICAL/C1202-12.htm (Accessed date: 16/1/2020).

[42] Y. Liu, F.J. Presuel-Moreno, M.A. Paredes, Determination of chloride diffusion coefficients in concrete by electrical resistivity method, American Concrete Institute Materials Journal 112 (5) (2015), https://doi.org/10.14359/51687777.

[43] J.K. Su, C.C. Yang, W.B. Wu, R. Huang, Effect of moisture content on concrete resistivity measurement, J. Chin. Inst. Eng. 25 (1) (2002) 117–122, https://doi.org/10.1080/02533839.2002.9670686.

[44] J.M. Sanchez Marquez, Influence of saturation and geometry on surface electrical resistivity measurements, Concordia University, 2015 Ph.D. Thesis, Website, (Accessed date: 16/1/2020) https://spectrum.library.concordia.ca/980090/1/Sanchez_M.A.Sc_F2015.pdf.

[45] A. Shahroodi, Development of test methods for assessment of concrete durability for use in performance-based specifications, University of Toronto, 2010 Ph.D. Thesis,Website, Website https://tspace.library.utoronto.ca/bitstream/1807/25796/5/shahroodi_ahmad_201011_MASc_thesis.pdf.

[46] E. Güneyisi, T. Özturan, M. Gesoğlu, A study on reinforcement corrosion and related properties of plain and blended cement concretes under different curing conditions, Cem. Concr. Compos. 27 (4) (2005) 449–461, https://doi.org/10.1016/j.cemconcomp.2004.05.006.

[47] A.A. Ramezanianpour, A. Pilvar, M. Mahdikhani, F. Moodi, Practical evaluation of relationship between concrete resistivity, water penetration, rapid chloride penetration and compressive strength, Constr. Build. Mater. 25 (5) (2011) 2472–2479, https://doi.org/10.1016/j.conbuildmat.2010.11.069.