Construction and Building Materials 348 (2022) 128566

Contents lists available at ScienceDirect

Construction and Building Materials

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

A machine learning method for predicting the chloride migration coefficient of concrete

Woubishet Zewdu Taffese *, Leonardo Espinosa-Leal

School of Research and Graduate Studies, Arcada University of Applied Sciences, Helsinki, Finland

ARTICLE INFO

Keywords:
XGBoost
Non-steady-migration coefficients
Machine learning
Concrete durability
Permeability
Chloride transport

ABSTRACT

This work adopts a state-of-the-art machine learning algorithm, XGBoost, to predict the chloride migration coefficient (D_{nssm}) of concrete. An extensive database of experimental data covering various concrete types is created by gathering from research projects and previously published studies. A total of four D_{nssm} models are developed depending on the number and type of input features. All models are verified with unseen data using four statistical performance indicators and compared to other five tree-based algorithms. The verification results confirm that the XGBoost model predicts the D_{nssm} with high accuracy. The model has the potential to replace cumbersome, time-consuming and resource-intensive laboratory testing.

1. Introduction

Chloride attack on reinforced concrete (RC) structures exposed to marine environments and deicing salts containing chloride is one of the most common threats to their durability. Usually, the highly alkaline environment of concrete builds a passive layer on the surface of reinforcement bars that prevents the bar from corrosion. Nonetheless, when the chloride concentration amount at the reinforcement bar reaches a certain level, depassivation (deterioration of the passive protection layer) occurs, causing corrosion and ultimately reducing the structure's safety, serviceability, and durability [1–7]. Understanding the chloride transport process is of the utmost importance to extend the durability and the service life of structures exposed to chloride-laden environments. On the other hand, the transport of chloride ions in concrete is a complicated chemical and physical process involving various transport mechanisms such as diffusion, capillary suction, and permeation [1,8,9]. This process is simplified by assuming that diffusion is the principal mechanism for chloride transport into the concrete medium. Fick's second law of diffusion, in which the diffusion coefficient is considered to be constant, is a universally applied mathematical model to determine the non-steady-state diffusion as presented by Eq. (1).

$$\frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial x^2} \tag{1}$$

where $C(x,t)$ is the chloride content at depth x and time t, D is the chloride diffusion coefficient.

Various laboratory test methods have been proposed to measure the chloride diffusion coefficient of concrete. As specified in ASTM C1556–11 [10] and NT Build 443 [11], the bulk diffusion tests are long-term experiments in which concrete specimens are exposed to a chloride solution for an extended period. On the other hand, such test methods are often not preferred in practice because they are laborious and time-consuming. One of the accelerated test methods for the chloride diffusion coefficient is the Nordic standard NT Build 492 [12], in which chlorides permeate the concrete at high rates due to the applied electric field. The diffusion coefficient determined by this method is referred to as the "non-steady-state migration coefficient" or D_{nssm} (also in this study) to distinguish it from bulk diffusion tests. Indeed, this migration coefficient cannot be directly compared to chloride diffusion coefficients obtained from other test methods. Though the test method of NT Build 492 provides rapid results, the test is typically performed after 28 days of the concrete production as it needs to be cured. The test also requires an experienced operator. It is thus difficult to experimentally evaluate the diffusion coefficient of concrete for each project due to the high related time and resources required. Therefore, it is vital to develop models that determine the chloride diffusion coefficients for the specific concrete by considering all influential parameters.

Significant efforts have been made in recent years to build phenomenological and physically-based chloride diffusion coefficient prediction models that consider factors describing concrete mix ingredients. For instance, Chidiac and Shafikhani [13] proposed a phenomenological model based on tortuosity factor, aggregate volume

* Corresponding author.
E-mail address: woubishet.taffese@arcada.fi (W.Z. Taffese).

https://doi.org/10.1016/j.conbuildmat.2022.128566
Received 28 February 2022; Received in revised form 1 July 2022; Accepted 22 July 2022
Available online 9 August 2022

fraction, porosity, chloride diffusivity of cement paste, compressive strength, the content of cement and supplementary cementitious material (SCM) to quantify the effective chloride diffusion in concrete. Riding et al. [14] proposed a model to estimate the apparent diffusion coefficient of concrete with different types of SCM. Several relationships in the model are used to calculate the diffusion coefficient. Bogas & Gomes [15] provide empirical expressions for calculating the diffusion coefficient as a function of the water-to-cement ratio (w/c). Sun et al. [16] used a multiscale method to calculate the effective diffusion coefficient of chloride ions in cementitious materials using a spherical *N*-layer inclusion model. They then devised a general equation for the effective medium that took into account the microstructure to estimate the value of the diffusion coefficient of the hardened cement paste. Audenaert et al. [9] proposed mathematical equations to estimate the non-steady-state migration coefficient in conventional and self-compacting concrete (SCC) based on the reference migration coefficient and the corresponding age factor related to capillary porosity. A thorough examination of relevant models can be seen in [17].

The proposed models consider only a few factors, leaving out some crucial ones that characterize concrete microstructure. The accuracy and applicability of these models vary significantly under different settings as they are constructed based on several model assumptions and different experimental databases. In addition, a rapid increase in the use of SCMs and chemical admixtures are other factors limiting the applicability of the models for efficiently predicting the diffusion coefficient. Hence, developing methods that consider all the governing parameters are essential. Certainly, developing an advanced chloride diffusion coefficient model that addresses all the influential parameters is a challenging task because the permeability property of concrete is a function of numerous parameters that are hard to represent mathematically without considering several assumptions. To counter these issues, developing a diffusion coefficient model using state-of-the-art machine learning algorithms could be a better alternative since they are powerful in solving complex problems involving large numbers of variables without making any assumptions. Machine learning has been instrumental in improving the productivity of many services and industries. Although its use in the construction industry is still in its infancy, its use has increased in recent years to address several issues such as geotechnics [18–20], and concrete durability [21–25]. A comprehensive overview of the use of machine learning in concrete durability can be found in [26].

This work offers a threefold contribution. These are i) the development of machine learning-based prediction models for the chloride migration coefficient of concrete by adopting the state-of-the-art algorithm, XGBoost, ii) the use of employing a wide range of concrete mixes, and iii) the investigation of the influence of fresh and hardened concrete tests on the prediction of chloride migration coefficients.

2. Related work

Machine learning approaches have been embraced for solving complex civil engineering problems throughout the last few decades. They have a significant potential to capture interdependencies between input and output datasets that are nonlinear, unspecified, or complicated to devise. It has been used in recent years to solve complex concrete durability problems, focusing on chloride transport. It includes the prediction of surface chloride concentration, chloride penetration, chloride diffusion coefficient and classification of chloride migration resistance of concrete. For instance, studies performed by Cai et al. [27] and Ahmad et al. [28] demonstrate the applicability of machine learning algorithms to predict surface chloride concentration. Both works applied the same data source to train and test the models. The database obtained many observations of case concrete elements exposed to marine environments that comprise surface chloride concentration along with concrete mix ingredients, environmental conditions, and exposure time. In terms of models, Cai et al. [27] developed five standalone machine learning models, including Gaussian process regression (GPR), multilayer perceptron artificial neural network (MLP-ANN), linear regression (LR), support vector machine (SVM), and random forests (RF), as well as an ensemble weighted voting-based model (RF + MLP + SVM), and then compare their prediction results. The authors claimed that the ensemble model achieves superior prediction accuracy than the standalone. In contrast, Ahmad et al. [28] utilized gene expression programming (GEP), decision tree (DT), and artificial neural network (ANN). The authors reported that the GEP model is the most accurate compared to the others.

Mohamed et al. [29] proposed ANN to estimate the degree of chloride penetration in SCC comprising different amounts of SCMs, including fly ash (FA), silica fume (SF) and ground-granulated blast-furnace slag (GGBS). The authors reported that the developed model made predictions with an accuracy of 96.6 %. Najimi et al. [30] proposed a hybrid of feed-forward artificial neural networks with an artificial bee colony algorithm (FF-ABC) to assess chloride penetration in SCCs with different mixture fractions. The authors used the water-to-binder ratio (w/b) as an input parameter along with others describing the number of binders, aggregates, and chemical admixtures. They compared the model's performance to the LR, genetic algorithm (GA), and particle swarm optimization (PSO)-based models. Based on the statistical measure of performance, the authors claimed that the FF-ABC model outperforms the other models. A study conducted by Kumar et al. [31] claimed that multivariate adaptive regression splines (MARS) and minimax probability machine regression (MPMR) are promising algorithms to estimate chloride penetration into concrete. The experimental data used to train and test the models are SCCs that include SCMs, namely FA and SF.

To predict chloride diffusion in cement mortar, Hoang et al. [32] proposed MARS and multi-gene genetic programming (MGGP). The authors reported the superior performance of the models after comparing their performance to ANN and least squares support vector regression (LSSVR). They may assist in discovering significant parameters that control chloride ion diffusion in cement mortar. Hodhod and Ahmed [33] adopted the ANN algorithm to assess the chloride diffusivity of high-performance concrete (HPC). Cement content, w/b, FA or GGBFS content, and curing age were the four input parameters used in the ANN model. The authors reported that the model predicted chloride diffusion coefficient with high accuracy based on the performance evaluation. The use of backpropagation (BP) neural network and the PSO on the BP neural network to predict the chloride penetration in concrete is presented in [34]. Various types of mineral admixtures (such as GGBS, FA, and SF) were used to produce the experimental concrete specimens. The authors claimed that the PSO-BP neural network provides a better estimate than the BP neural network.

A study by Delgado et al. [35] adopted ANN to determine the depth of chloride penetration and the diffusion coefficient of concrete specimens under conditions of drying–wetting cycles. Type of cement, w/c, type of mineral additives, curing age, and the number of drying-wetting cycles as predictors. The authors reported that the developed model adequately predicts both parameters.

Marks et al. [33] classified the degree of resistance to chloride penetration in concrete modified with high calcium fly ash (HCFA) using 20 machine learning algorithms based on migration coefficient values. There are three types of models used: i) Bayesian classifiers (three algorithms), ii) tree classifiers (nine algorithms), and iii) rule classifiers (eight algorithms). The J48 algorithm, a tree-based classifier, provided the highest accuracy of all. In another relevant study performed by Marks et al. [36], the J48 algorithm was used to classify the chloride migration resistance of concrete modified with circulating fluidized bed combustion (CFBC). The author of the paper claimed that the algorithms used were suitable for classification. Table 1 provides additional information on the previously presented work.

Table 1
Machine learning models proposed to predict factors related to chloride transport in concrete.

Work	Predicted feature	Involved test	Concrete type	Exposure environment		Machine learning algorithm	Input category			No. of input features	No. of observations	Problem type
				Lab	Field		Mix ingredients	Fresh property	Hardened property			
Cai et al. [27]	Surface chloride concentration	N/A	Various type	X	✓	LR, GPR, SVM, MLP-ANN, RF, Ensemble (RF + MLP + SVM)	✓	X	X	12	642	Regression
Ahmad et al. [28]	Surface chloride concentration	N/A	Various type	X	✓	DT, GEP, ANN	✓	X	X	12	642	Regression
Mohamed et al. [29]	Chloride penetration level	ASTM C1202	SCC	✓	X	ANN	✓	X	X	13	72	Regression
Najimi et al. [30]	Chloride permeability	ASTM C1202	SCC	✓	X	FF-ABC, LR, GA, PSO	✓	X	X	6	72	Regression
Kumar et al. [31]	Chloride permeability	ASTM C1202	SCC	✓	X	MARS, MPMR	✓	X	X	3	360	Regression
Hoang et al. [32]	Chloride diffusion coefficient	AASHTO T260-97	Mortar	✓	X	MGGP, MARS, LSSVM, LM-ANN	X	X	X	4	132	Regression
Hodhod and Ahmed [33]	Chloride diffusion coefficient	ASTM C1202	HPC	✓	X	ANN	✓	X	X	4	300	Regression
Yao et al. [34]	Chloride diffusion coefficient	ASTM C1202	NWC	✓	X	PSO-BP, BP	✓	X	X	8	120	Regression
Delgado et al. [35]	Chloride penetration depth and chloride diffusion coefficient	ASTM C1202	NWC	✓	X	ANN	✓	X	X	5	243	Regression
Marks et al. [37]	Chloride resistance level	NT Build 492	NWC contains HCFA	✓	X	J48, 19 other types of classifiers	✓	X	X	4	56	Classification
Marks et al. [36]	Chloride resistance level	NT Build 492	NWC contains CFBC	✓	X	AQ21, J48	✓	✓	✓	6	15	Classification

3. Research significance

The machine learning methods utilized to solve various problems related to chloride transport have yielded promising results, but much work remains to be done. The limitations of the earlier works can be outlined as follows. First, most experimental data on chloride diffusion are not based on NT Build 492. Second, the experimental data for each work is limited to a specific concrete type, such as SCC, HPC, normal weight concrete (NWC), and cement mortar. Third, some studies have not considered many parameters in the mix constitutes category that significantly controls or describe the chloride transport in concrete. Fourth, almost all studies failed to consider fresh and hardened tests as predictors. It is well known that the transport of chlorides into the concrete pores is governed by the concrete microstructure, which is immensely complex and strongly influenced by the nature of the mix components and proportions.

The previously proposed machine learning-based models may fit the experimental data considered in their work, but they may not be well suited to predicting the chloride permeation resistance of concrete using other experimental data. It is understandable that all potential influencing parameters and a comprehensive database must be considered to develop a reliable, universal, and robust model for predicting the chloride permeability property of different types of concrete. To this end, this work thoroughly evaluates the potential of the XGBoost algorithm to predict the chloride migration coefficient to address the above shortcomings. An extensive database comprised of 843 observations and 23 features acquired from numerous peer-reviewed articles and research projects is used for chloride migration coefficient model development. A detailed description of the data is provided in Section 5.1. Using such a large dataset is critical to developing a universal and robust model. Substantial optimization techniques are also applied in this work to enhance the performance of the models.

4. XGBoost algorithm

XGBoost stands for “eXtreme Gradient Boosting” and is designed using the general principles of gradient boosting, combining weak learners to a strong learner [38]. Gradient boosted trees are often formed in a sequential manner, gradually learning from data to improve prediction in each iteration. XGBoost, on the other hand, creates trees in parallel, and it improves prediction performance by managing model complexity and reducing overfitting through built-in regulation. Due to these distinct advantages, speed and performance, XGBoost has become the dominant machine learning algorithm for solving regression problems in several civil engineering applications, e.g. prediction of porosity [21], shear strength of concrete-to-concrete interface [39], and the

residual value of construction equipment [40].

The final strong XGBoost model $F()$ can be described by Eq. (2).

$$\widehat{y}_i = F(X_i) = \sum_{k=1}^{K} f_k(X_i), \tag{2}$$

where $\widehat{y}_i$ is the prediction for the $i-th$ sample; $f_k(.)$ is the $k-th$ weak learner of the strong model, that is the $k-th$ decision tree; K is the total number of the weak learners.

The objective function throughout the training process is presented by Eq. (3).

$$Obj(\theta) = \sum_{i=1}^{N} l(\widehat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where $l(\cdot)$ is the loss function; y_i is the actual value of the $i-th$ sample; $\Omega(\cdot)$ is the regularization term.

Eq. (4) shows the prediction for the $i-th$ sample at the $k-th$ iteration.

$$\widehat{y}_i^{k} = \widehat{y}_i^{k-1} + f_k(X_i) \tag{4}$$

where $\widehat{y}_i^{k-1}$ is the prediction from the preceding cumulative model after $(k-1)-th$ iteration The objective function can therefore be rewritten as shown in Eq. (5).

$$Obj = \sum_{i=1}^{N} l\left(\widehat{y}_i^{k-1} + f_k(X_i), y_i\right) + \Omega(f_k) \tag{5}$$

As the XGBoost uses a second-order Taylor approximation of the loss function, the objective function can be stated roughly as shown in Eq. (6).

$$Obj \cong \sum_{i=1}^{N} l\left(\widehat{y}_i^{k-1}, y_i\right) + g_i f_k(X_i) + \frac{1}{2} h_i f_k^2(X_i) + \Omega(f_k) \tag{6}$$

where $g_i = \partial \widehat{y}^{(k-1)} l\left(\widehat{y}_i^{k-1}, y_i\right)$, and.$h_i = \partial^2 \widehat{y}^{(k-1)} l\left(\widehat{y}_i^{k-1}, y_i\right)$

The decision tree for a weak learner can be expressed by Eq. (7).

$$f_k(X_i) = w_{q(X_i)} \tag{7}$$

where q and w represent the tree's structure and leaf weights, respectively.

The regularization term $\Omega(f_k)$ can be written as shown in Eq. (8).

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{8}$$

where T is the total number of leaves, γ and λ represent the penalty coefficients.

The objective function of XGBoost can be further simplified as presented in Eq. (9).

$$\begin{aligned} Obj &\cong \sum_{i=1}^{N}\left[g_i f_k(X_i) + \frac{1}{2} h_i f_k^2(X_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T. \end{aligned} \tag{9}$$

As a result, the decision tree's optimal weights w_j^* and the objective function's optimal value can be computed by Eq. (10) and Eq. (11), respectively.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{10}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{11}$$

5. Materials and methods

This section focuses on the experimental data used as well as the overall model development process, from data preprocessing to model training and evaluation. It starts with the presentation of the experimental dataset and then goes into detail about the model development process.

5.1. Experimental dataset

Non-steady-state migration coefficients (D_{nssm}) of various types of concrete performed according to NT Build 492 are considered in this study. The testing methodology of NT Build 492 is primarily based on the chloride ions migration into previously vacuum saturated concrete specimens in anolyte solution (0.3 M NaOH) by applying an external electrical voltage (30 V DC) that is adjusted between 10 and 60 V according to the electrical current in the scheme. The applied electrical potential forces the chloride ions from the 10 % NaCl solution (catholyte) to migrate into the concrete specimens. The test configuration of this method is depicted in Fig. 1. After a certain period, the specimens are split axially and silver nitrate solution ($AgNO_3$) is sprayed onto the freshly fractured surface, which reacts to give white insoluble silver chloride when it comes into contact with chloride ions. This allows for the measurement of chloride penetration depths across the split surface at 10 mm intervals, yielding 5 to 7 valid depth readings [12,41]. The average of chloride depths is taken as the representative value. Once the chloride depth is determined, the chloride migration coefficient is then calculated using Eq. (12) [12].

$$D_{nssm} = \frac{0.0239(273+T)L}{(U-2)t}\left(x_d - 0.0238\sqrt{\frac{(273+T)L.x_d}{U-2}}\right) \tag{12}$$

where D_{nssm} is the non-steady-state migration coefficient, x10^{-12} m^2/s; U is the absolute value of the applied voltage (V); T is the average initial and final temperatures in the anolyte solution (°C); L is the specimen's thickness (mm); x_d is the average penetration depths value (mm); t is the test duration (h).

A comprehensive database containing 843 experiments that examine the non-steady-state migration coefficients (D_{nssm}) of distinct concrete types is formed by collecting from: i) research projects (LIFECON and Finnish DuraInt-project [43]), and ii) internationally published journal articles by accessing Web of Science and Scopus databases [1,7–9,36,37,15,44,53–57,45–52]. The database retained information concerning the concrete mix, its fresh and hardened properties. The concrete mix comprises eight features describing the ingredients type and proportion. These are w/b, contents of binders: cement, slag, FA, SF, and lime filler (unit of kg/m^3), amount of fine, coarse, and total aggregates (unit of kg/m^3), contents of chemical admixtures: plasticizers, superplasticizers, and air-entraining agents (AEA) in (% by binder wt.). The fresh and hardened properties describe different properties of the concrete specimens. The fresh concrete properties comprise tests of slump and slump flow (unit of mm) to describe the workability of the concrete; air content (in %); as well as fresh and dry density (unit of kg/m^3). The hardened concrete property test includes test results of comprehensive strength (unit of MPa) and non-steady-state migration coefficients (unit of x10^{-12} m^2/s) performed at different maturity ages.

As the database is built based on experimental data carried out in different parts of the world, it consists of different types of cement specified in various standards. Hence, for the sake of consistency, all the cement types are translated as per European Standard EN 197–1 [58]. It defines 27 distinct common types of cement that can be grouped into five (CEM I, II, III, IV, and V). In the database, a total of 15 varieties of cement from the four basic cement types are included. These are Portland cement (CEM I), Portland-slag cement (CEM II/A-S and CEM II/B-S), Portland-silica fume cement (CEM II/A-D), Portland-fly ash cement (CEM II/A-V, and CEM II/B-V), Portland-limestone cement (CEM II/A-L,

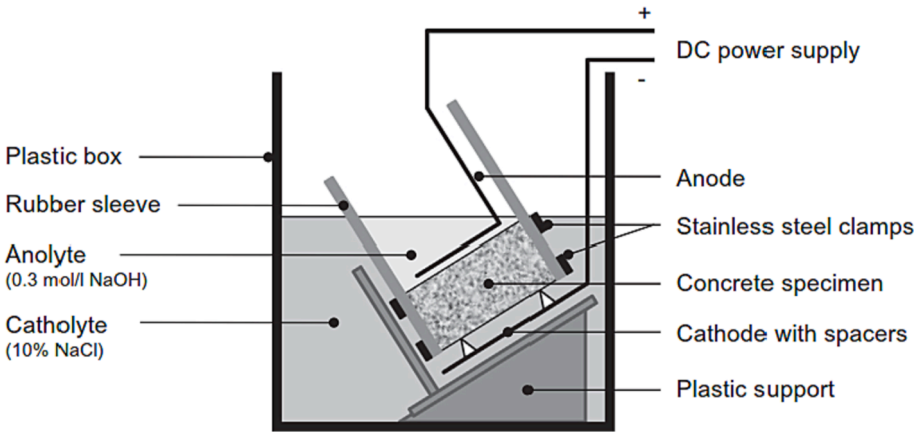

Fig. 1. NT Build 492 test set-up [42].

CEM II/B-L, and CEM II/A-LL), Portland-composite cement (CEM II/A-M and CEM II/B-M), blast furnace cement (CEM III/A and CEM III/B), and pozzolanic cement (CEM IV/A and CEM IV/B). Approximately 57 % of the experiments use supplementary cementitious materials of slag, FA, SF, or lime filler. The w/b varies between 0.19 and 0.65. Eight types of chemical admixtures: one (magnesium lignosulfonate) as a plasticizer, four (naphthalene, polycarboxylate ether, melamine sulfonate, and lignosulfonate) as a superplasticizer, and three (fatty acid soap, vinsol resin, and synthetic surfactants) as AEA are used. Different varieties of concrete are included in the database, including regular strength, lightweight, high-strength, high-performance, and self-consolidating concrete. Table 2 describes the data in greater detail.

5.2. Model development

This section describes the model development approach for predicting D_{nssm}. Fig. 2 illustrates the flow of the model development process. Obtaining the dataset, including concrete ingredients and proportions, fresh and hardened concrete properties, and non-steady-migration coefficients is the first activity. Then data preprocessing which is the most important step in the development of a machine learning model is followed as data from real-world scenarios is generally noisy, contains missing values, and may even be in an unusable format that cannot be directly employed for use in a machine learning model. To make the data appropriate for a machine learning model, a variety of activities are typically performed during data preprocessing, including outlier detection and treating, data encoding, data normalization,

Table 2
Description of features employed in the raw dataset.

Feature category	No.	Feature subcategory		Description	Unit	Min value	Max value
Concrete mix ingredients	1	Cement types	CEM I	CEM I	–	–	–
			CEM II	CEM II/A-S, CEM II/B-S, CEM II/A-D, CEM II/A-V, CEM II/B-V, CEM II/A-L, CEM II/B-L, CEM II/A-LL, CEM II/A-M, CEM II/B-M	–	–	–
			CEM III	CEM III/A, CEM III/B	–	–	–
			CEM IV	CEM IV/A, CEM IV/B	–	–	–
	2	Water content			[kg/m^3]	8.46	1049.39
	3	Cement content			[kg/m^3]	13.02	2384.97
	4	Mineral admixtures content		Slag	[kg/m^3]	0.00	1284.44
	5			Fly ash	[kg/m^3]	0.00	735.00
	6			Silica fume	[kg/m3]	0.00	468.50
	7			Lime filler	[kg/m^3]	0.00	350.00
	8	Water-to-binder ratio			–	0.19	0.65
	9	Aggregates content		Fine aggregate	[kg/m^3]	27.53	1574.10
	10			Coarse aggregate	[kg/m^3]	0.00	1240.00
	11			Total aggregate	[kg/m^3]	54.04	2097.00
	12	Chemical admixtures content		Plasticizer	[% by binder wt.]	0.00	0.89
	13			Superplasticizer	[% by binder wt.]	0.00	4.17
	14			Air-entraining agent	[% by binder wt.]	0.00	6.50
Fresh concrete properties	15	Basic properties		Slump	[mm]	4.00	250.00
	16			Spread	[mm]	9.53	598.00
	17			Air content	[%]	1.10	8.00
	18			Fresh density	[kg/m^3]	1364.00	2609.00
	19			Dry density	[kg/m^3]	1225.00	2427.03
Hardened concrete properties	20	Mechanical properties		Compressive strength	[MPa]	16.90	483.00
	21			Concrete age at compressive strength test	[days]	7	180
	22	Migration properties		Concrete age at migration test	[days]	3	365
	23			Migration coefficient (D_{nssm})	[x10^{-12} m^2/s]	0.15	133.60

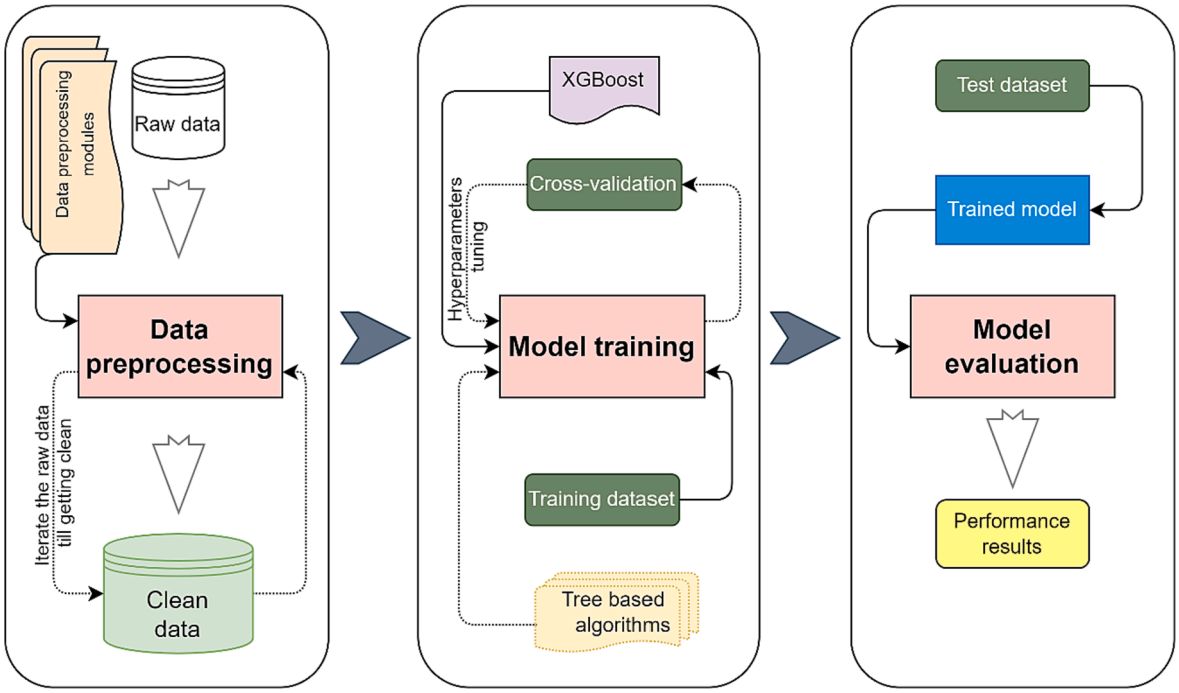

Fig. 2. Model development process of chloride migration coefficient prediction.

feature engineering, and splitting the dataset into a training and test set. The next step is selecting the right algorithms and training the model using the training dataset. The process of model training will be iterated until the best cross-validation results are obtained by optimizing the hyperparameters. The models' performance is then assessed using a test set that is previously unseen by the model. The following sections cover all major activities involved in the model developing process.

Four models are created based on the input features. The goal is to investigate the significance of fresh and hardened concrete properties in predicting the chloride migration coefficients. Table 3 shows the classifications of the four models. Model I utilize all the 14 input features presented in Table 1 under the category of concrete mix ingredients. Model II employ input features representing only concrete mix ingredients and fresh concrete properties. Model III consider 16 features presented under the category of concrete mix ingredients and hardened concrete properties, while Model IV utilize all the features from Table 1 except the migration coefficient, D_{nssm}. Indeed, after transforming the raw data into a useful and efficient format in the data preprocessing phase, the number of features used to train the model could be changed.

5.3. Data preprocessing

Data preprocessing is a critical step in the development of any machine learning-based models. Missing data processing, detecting and treating outliers, data encoding, data normalization, and data partitioning are among the frequently applied tasks under data preprocessing. All the data preprocessing tasks applied in this work are detailed in the following subsections.

Table 3
Classification of the four models.

Models	Input features category	No. of input features	Number of observations
Model I	Concrete mix ingredients	14	843
Model II	Concrete mix ingredients Fresh concrete properties	19	843
Model III	Concrete mix ingredients Hardened concrete properties	16	843
Model IV	Concrete mix ingredients Fresh concrete properties Hardened concrete properties	22	843

5.3.1. Missing data processing

One of the most essential components in improving the prediction accuracy of any data-driven model is the quality of the input data. Missing data is defined as values that do not exist in the specified dataset for some features. It diminishes the predictive ability of machine learning models as well cause the model to be biased, and it is one of a problem that affects almost all scientific fields. As a result, missing data must be addressed first. It can be dealt with in a number of methods, including i) omitting observations with any missing values, ii) relying on the learning algorithm to deal with missing values during training, and iii) imputing all missing values prior to training. In this work, all the missed observation are discarded.

5.3.2. Detecting and treating outliers

Outliers are unusual observations that are extremely distant from the rest of the population. Any data-driven model development process should include detecting and treating outliers since the model's performance depends on data quality. Indeed, not all machine learning algorithms are sensitive to outliers, but the adopted algorithm (XGBoost) is extremely sensitive to outliers. Outlier detection approaches that focus on each variable separately, recognizing extreme observations based on the observed univariate distribution, are some of the most common. However, this approach does not identify outlying observations given relationships between two or more features. Hence, the adoption of the multivariate outlier detection method is essential to identify situations in which two or more features have an uncommon combination of scores. There are two types of multivariate outlier detection techniques [59]: i) distance-based approaches and ii) lower-dimensional projection-based methods. The Mahalanobis Distance (MD) is a widely used distance measure in multivariate space that takes into account the data's mean and covariance, and returns bigger distances for observations that deviate from the mean in directions with lower covariance. In this work this approach is employed to detect

multivariant outliers. In fact, the MD must be compared to a cut-off value derived from the chi-square distribution in order to detect multivariate outliers. The Mahalanobis distance between a two of objects X_A and X_B is defined by Eq. (13) [60].

$$d = \left[(X_B - X_A)^T.C^{-1}.(X_B - X_A)\right]^{0.5} \quad (13)$$

where C is the covariance matrix of the sample.

The Mahalanobis distance can also be calculated from each observation to the data center, as shown in Eq. (14) [60].

$$d_i = \left[(X_i - \overline{X})^T.C^{-1}.(X_i - \overline{X})\right]^{0.5}, \quad (14)$$

where X_i is an object vector, and $\overline{X}$ is an arithmetic mean vector.

5.3.3. Data encoding

Many machine learning models need their input variables to be numeric, and thus any categorical features need to be transformed. In the employed dataset, the feature "Cement type" comprises nominal variables (nonnumeric and descriptive data types). One-hot encoding is the most commonly applied to translate nominal variables to numeric, in order to improve the performance of the algorithm. It converts categorical data to a binary vector format. In other words, every distinct value in a column results in a new column. Based on whether the value matches the column header, this column is represented as 1 or 0 s. Fig. 3 provides an example of one-hot encoding in the case of Model III. As can be seen, there are six categories for the feature Cement type in Model III, which requires six binary variables. Using one-hot encoding, the categorical variables, CEM I, CEM II/A-D, CEM II/A-S, CEM II/B-S, CEM II/B-V, and CEM III/A are denoted as binary variables [1,0,0,0,0,0], [0,1,0,0,0,0], [0,0,1,0,0,0], [0,0,0,1,0,0], [0,0,0,0,1,0], and [0,0,0,0,0,1].

5.3.4. Feature selection

Feature selection is the process of extracting the most important features from a dataset. It can support the performance of a machine learning model. Feature selection methods are classified into three types: filter, wrapper, and embedded [61]. The embedded method is used in this work because it combines the advantages of the filter and wrapper methods in terms of low computational effort and sufficient accuracy. An embedded method based on a random forest algorithm is utilized to select relevant features from the dataset. Without making any assumptions about the data, this algorithm measures the importance of a feature as the averaged impurity decrease deduced from all decision trees in the forest. As an example, Fig. 4 shows the feature importance measured by a random forest in the case of Model I. After removing the missing values, the correlation coefficients between all possible features are presented in Fig. 5 to help understand the dependency between features. The feature importance measures presented in Fig. 4 added up to one. It is evident that the features w/b and coarse aggregate have the greatest predictive power. These two features accounted for 57 % of the model's predictive power, followed by water (13 %), and binders (11.7 %). Indeed, two-thirds of the predictive power observed in binders is attributed to supplementary cementitious materials (slag, fly ash, and silica fume). Although water is a very important feature, w/b has already encoded its information. This fact is confirmed by a strong correlation coefficient (0.57) between water and w/b as shown in Fig. 5. The same phenomenon is also observed in the case of total aggregate. This feature has a high correlation coefficient, 0.72, with coarse aggregate. This means that the total aggregate feature is redundant since it is encoded by a coarse aggregate and should not be used in model training. Cement types have the lowest predictive power of all the features. They contributed only 1.5 % to the predictive power of the random forest model. The number of cement types shown in Fig. 4 is only seven, although the raw data set presented in Table 2 contained 15 cement types. This is because observations with missing values were removed during missing data preprocessing, reducing the total number of cement types in Model I to seven. Air-entraining, plasticizer, and lime filler have no feature importance measure and are not presented in Fig. 4 because these features only have zero values after all missing values are removed from the dataset, as depicted in Fig. 5. Based on all these facts only features (w/b, cement, slag, fly ash, silica fume, fine aggregate, coarse aggregate, superplasticizer, migration test age, cement types) were selected as D_{nssm} predictors for Model I.

5.3.5. Data partitioning

Typically, training/test partitioning involves partitioning the data into a training and a test set in a specific ratio. The training set is used to train the model, while the test set is used to evaluate the fitted model's predictive performance against data it has never seen before. In this work, the data are randomly divided into two parts: 80 % and 20 %. 80 percent of the data is used as a training set and 20 percent of the data is used as a test set to develop all four models.

5.4. Model training and evaluation

After preprocessing the data, all the four models are trained on the corresponding training sets. Table 4 shows a description of the data after preprocessing. Even though the raw data set contains 843 observations with 23 features (including the migration coefficient), the number of features and observations used to train the models is significantly reduced after data preprocessing. For instance, the number of observations in Models I, II, III and IV are 134, 131, 176, and 96, respectively. Descriptive statistics of the pre-processed data for the four models are presented in Table 5. The hyperparameters of the XGBoost algorithm are then optimized using a grid search method to define a high-performance model for predicting the chloride migration coefficient. This method performs an exhaustive search through a manually defined subset of a learning algorithm's hyperparameter space. A grid search algorithm is typically guided by a cross-validation or hold-out performance metrics. In this work, the K-fold cross-validation method is applied. In this method, the training set is randomly divided into K subgroups of roughly equal size. Each of the K subsets serves as a validation set to assess the model's performance, while the remaining (K − 1) subsets serve as a training set. In total, K models are fitted, and K validation statistics are obtained. The score from the K-folds is averaged to determine the overall performance of the model.

The grid search method, combined with the 5-fold cross-validation technique, was used to find the best hyperparameters for all models. The following hyperparameters were considered along with their grid search ranges: (i) Number of gradient boosted trees ('n_estimators'): [20, 50, 100, 300, 500], (ii) Maximum tree depth for base learners ('max_depth'): [2,4,6,8,10], (iii) Boosting learning rate

Cement type
CEM I
CEM II/A-D
CEM II/A-S
CEM II/B-S
CEM II/B-V
CEM III/A

One-hot encoding

CEM I	CEM II/A-D	CEM II/A-S	CEM II/B-S	CEM II/B-V	CEM III/A
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1

Fig. 3. One-hot encoding for the feature "Cement type" in the case of Model III.

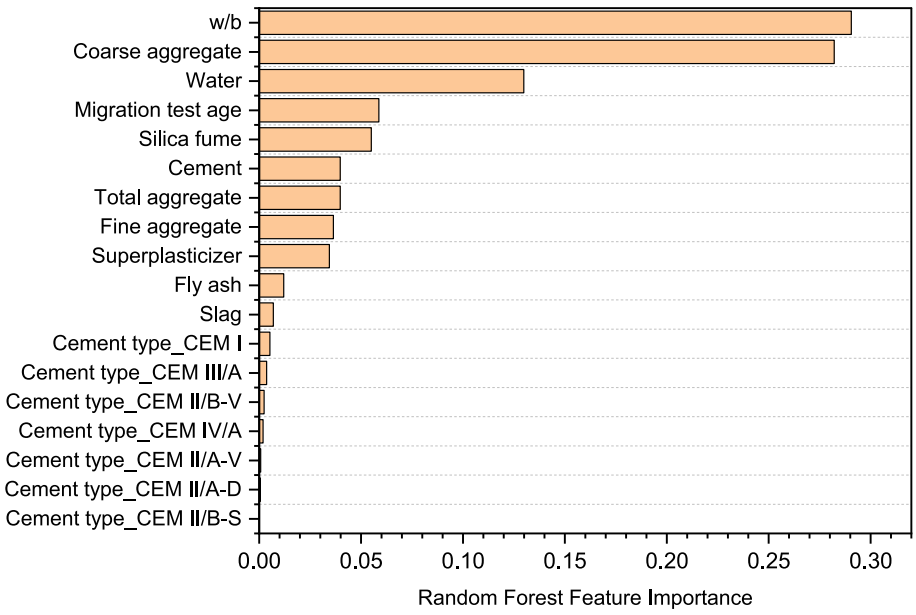

Fig. 4. The measure of feature importance in the case of Model I.

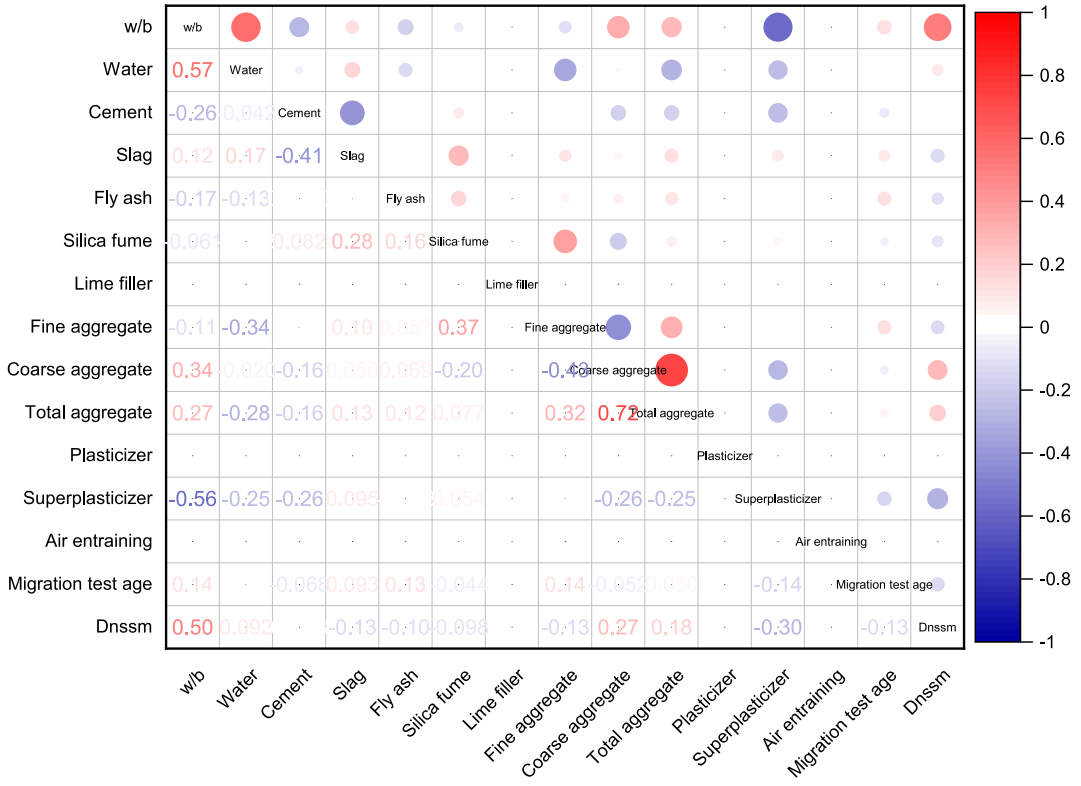

Fig. 5. Heat map of the correlation matrix of features describing concrete mix ingredients and migration coefficient.

('learning_rate'): [0.001, 0.01, 0.1, 0.3], (iv) Booster ('booster'): ['gbtree'], and (v) Minimum loss reduction required to make a further partition on a leaf node of the tree ('gamma'): [0.0001, 0.001, 0.01]. The other hyperparameters have been left at their default settings. The created grid (configuration of 5 'n_estimators', 5 'max_depth', 4 'learning_rate', 1 'booster', and 3 'gamma') has a total of 300 configurations, and each combination is evaluated using 5-fold cross-validation, resulting in the construction of 1500 models. Table 6 depicts the optimized hyperparameters. The hyperparameters that led to the best prediction accuracy differ between the models. For instance, in the case of Model I, the number of gradient-boosted trees is 300 while it is 100 in the other models. With the exception of Model I, the 'gamma' value for all models that produced optimal accuracy was 0.01. The optimized values for the boosting learning rate are 0.1 or 0.3.

Once the best hyperparameters have been determined for each model, the next step is to utilize the optimal values in construction of the XGBoost model to predict D_{nssm}. A total of four models (Model I, Model II, Model III, and Model IV) are being developed. As shown in Table 4,

Table 4
Details of the features considered in each model.

Models	No. of input feature types	Description of input features	Number of observations
Model I	10	Basic features: (w/b, Cement, Slag, Fly ash, Silica fume, Fine aggregate, Coarse aggregate, Superplasticizer, Migration test age, Cement types).	134
Model II	11	Basic features (as in Model I) and Fresh density	131
Model III	12	Basic features (as in Model I), Compressive strength test age and Compressive strength.	176
Model IV	13	Basic features (as in Model I), Fresh density, Compressive strength test age and Compressive strength.	91

Remark: All of the input features have the same units as described in Table 2.

Model I uses ten types of input features (predictors) from the category of concrete mix ingredients. Model II employs the same features as Model I plus an additional feature (fresh density) describing the fresh property of concrete. Model III includes the same features as Model I, plus two new ones: compressive strength test age and compressive strength. Model IV employs all of the feature types found in the three models.

The ultimate goal of any predictive model is to perform well on previously unseen data. The performance of the models on new data is evaluated using statistical metrics of mean-square error (MSE), root-mean-square error (RMSE), mean-absolute error (MAE) and coefficient of determination (R^2) on the training and test sets. MSE is calculated by averaging the squared difference between the actual and predicted values, as given by Eq. (15). It is the most widely used loss function for regression models. RMSE is equal to the square root of MSE, Eq. (16). It is sometimes preferred over MSE because MSE values are harder to understand due to the squaring effect. This is especially true when the target represents values in units of measurement. The MAE is an average of the absolute errors (the difference between the actual and the predicted value), as in Eq. (17) and is measured in the same units as the target feature. MAE is also known as the absolute loss. R^2 is the variance fraction of the response feature determined by the regression model. It is also considered the standardized version of MSE due to the improved interpretability of the model's performance. The value of R^2 is calculated using Eq. (18).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{MSE}{Var(y)} \tag{18}$$

where n is the number of observations, y_i is the actual target value, $\hat{y}_i$ is the predicted output value, $\bar{y}$ the mean value of the actual target, Var is the target variable variance.

6. Results and discussion

This section presents and discusses the performance of all four models developed. All models are built using the XGBoost algorithm with the aim of estimating the chloride migration coefficient of concrete. The main difference between the models is the number and type of input features. Model I contains only features that describe the ingredients of the concrete mix. The other models also contain information about the ingredients of the concrete mix, but add one or more features. Model II adds a fresh density input feature to characterize the fresh property of concrete. On the other hand, Model III adds two features that describe the hardened property of concrete. These are the compressive strength and age of the concrete at the time of the testing. Model IV includes all of the features described in the other three models.

Fig. 6 shows regression plots comparing the actual and predicted D_{nssm} values of all models during model training. The regression plots also show the corresponding R-squared scores for all models. It can be perceived from Fig. 6 that all models score close to one, confirming that these models fitted the data very well during the training phase. It can also be seen from Fig. 6 that Model I, in contrast to the other modes, has few observations of D_{nssm} values well above $28x10^{-12}$ m^2/s. Including such limited extreme values results in slightly lower performance.

The training residuals, the discrepancy between the actual and predicted D_{nssm}, of the developed four models are computed, and visualized using a boxplot with distribution curves as shown in Fig. 7. The median of the residuals is represented by a line in the box that spans the middle fifty percent (25th to 75th percentiles) of the observed values. Each whisker starts at the lowest value and progresses to the highest value. The mean is represented by a small square inside the box. Outliers are points (denoted here by a circled times) that are more than 1.5 times but less than 3 times the interquartile range (IQR) above the third quartile or below the first quartile. Any points greater than 3*IQR are referred to as extreme values, denoted with an asterisk.

The mean of the residuals for all models, as shown in Fig. 7, is almost in the center of the box and is symmetrically distributed about the mean, indicating a normal distribution. It can also be seen that the standard deviations of the models that measure how to spread out a normally distributed distribution vary. The smaller the standard deviation, the steeper the bell curve. Models II and IV have lower standard deviations, followed by Model III. Model I has the highest standard deviation, indicating that it is relatively less accurate than the other three models. In general, the residuals boxplot demonstrates that all models learn the complex nonlinear relationship of the input features very well to predict the chloride migration coefficient.

The validity of any machine learning model must be assessed using a test set derived from the original data but not included in the training set. This is because the models can contain errors due to high bias and variance during training. High bias can lead to underfitting by causing the algorithm to miss the relevant connection between the input and the target features. High variance can lead to overfitting, causing the algorithm to model the random noise in the training dataset rather than the expected outputs [62]. All models are validated with the test set. The actual and predicted D_{nssm} are shown in Fig. 8 as regression plots with the corresponding R^2 values. It can be observed that the R^2 is well above 0.80 for all models, confirming the high performance of the models. Model IV appears to have the best learning performance with (R^2 = 0.963), followed by Model II (R^2 = 0.888), Model III (R^2 = 0.865), and Model I (R^2 = 0.830). It is also worth noting that all models have a slight tendency to overlook D_{nssm} when it is large. This is because the dataset had a limited set of large chloride migration coefficients. For example, D_{nssm} values greater than 16×10^{-12} m^2/s account for only about 15 % of all observations.

The models' testing residuals along with their corresponding training residuals are shown in Fig. 9 as boxplots with distribution curves for easy comparison. Clearly, the Model IV test residuals are small, and their mean lies in the center of the box and is symmetrically distributed around it. Unlike the other models, the residuals distribution of Model IV is bell shaped, indicating that it has a normal probability distribution. Model II has significantly more extreme values than any other model. Model I has the broadest distribution, meaning it has the highest standard deviation and is less accurate than the other three models. Indeed, this is to be expected since Model I's training performance is the worst with the widest spread. Despite the differences in the residual

Table 5
Descriptive statistics of the dataset for each model after preprocessing.

Model I

	w/b	Cement	Slag	Fly ash	Silica fume	Fine aggregate	Coarse aggregate	Superplasticizer	Migration test age	CEM I	CEM II/A-D	CEM II/A-V	CEM II/B-S	CEM II/B-V	CEM III/A	CEM IV/A	D_{nssm}
count	134	134	134	134	134	134	134	134	134	134	134	134	134	134	134	134	134
mean	0.42	387.49	4.19	22.2	7.82	807.58	835.34	0.34	45.35	0.91	0.01	0.01	0.01	0.01	0.03	0.02	9.7
std	0.07	70.16	24.25	53.82	15.65	192.54	284.62	0.37	32.71	0.29	0.09	0.09	0.12	0.09	0.17	0.15	7.68
min	0.3	225	0	0	0	517	266.35	0	3	0	0	0	0	0	0	0	0.74
25 %	0.36	341	0	0	0	681.75	607	0	28	1	0	0	0	0	0	0	5.38
50 %	0.4	391	0	0	0	743	934	0.18	28	1	0	0	0	0	0	0	7.6
75 %	0.45	450	0	0	0	1002.5	1065.06	0.67	90	1	0	0	0	0	0	0	10.55
max	0.6	525	170	216	60	1150	1240	1.01	182	1	1	1	1	1	1	1	50.05

Model II

	w/b	Cement	Slag	Fly ash	Silica fume	Fine aggregate	Coarse aggregate	Superplasticizer	Fresh density	Migration test age	CEM I	CEM II/A-D	CEM II/A-S	CEM II/B-S	CEM II/B-V	D_{nssm}
count	131	131	131	131	131	131	131	131	131	131	131	131	131	131	131	131
mean	0.35	316.71	52.59	7.86	2.21	764.39	872.5	0.18	2163.79	28.5	0.6	0.02	0.23	0.05	0.11	8.57
std	0.08	101.71	64.78	22.71	7.63	138.52	292.33	0.32	209.52	14.09	0.49	0.12	0.42	0.21	0.31	3.51
min	0.22	186	0	0	0	517	255	0	1534	3	0	0	0	0	0	3.4
25 %	0.29	234.34	0	0	0	693.54	573.5	0	1933	28	0	0	0	0	0	6.29
50 %	0.35	271.72	0	0	0	751.68	1037.05	0	2266.29	28	1	0	0	0	0	7.6
75 %	0.4	403.2	106.79	0	0	806.26	1088.37	0.2	2303.45	28	1	0	0	0	0	9.68
max	0.54	525	186.88	100	36	1073	1187.15	1.45	2402.77	182	1	1	1	1	1	19.42

Model III

	w/b	Cement	Slag	Fly ash	Silica fume	Fine aggregate	Coarse aggregate	Superplasticizer	Comp. str. test age	Compressive strength	Migration test age	CEM I	CEM II/A-D	CEM II/A-S	CEM II/A-V	CEM II/B-S	CEM II/B-V	CEM III/A	CEM IV/A	D_{nssm}
count	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176
mean	0.37	335.44	39.14	11.9	4.89	797.54	894.64	0.19	24.82	44.27	32.66	0.65	0.02	0.17	0.01	0.03	0.1	0.01	0.02	9.16
std	0.09	89.56	60.39	35.57	12.87	168.97	266.23	0.31	16.23	13.49	17.95	0.48	0.13	0.38	0.08	0.18	0.3	0.11	0.13	4.95
min	0.22	186.88	0	0	0	321	266.35	0	7	19.17	3	0	0	0	0	0	0	0	0	1.69
25 %	0.3	250.96	0	0	0	693.54	731	0	7	32.95	28	0	0	0	0	0	0	0	0	6.1
50 %	0.36	344.1	0	0	0	759.99	1033.78	0	28	43.43	28	1	0	0	0	0	0	0	0	7.63
75 %	0.45	400	87.21	0	0	915.25	1087.48	0.37	28	50.58	28	1	0	0	0	0	0	0	0	10.5
max	0.55	525	186.88	192	60	1150	1240	1.1	91	80	90	1	1	1	1	1	1	1	1	25.2

Model IV

	w/b	Cement	Slag	Fly ash	Silica fume	Fine aggregate	Coarse aggregate	Superplasticizer	Fresh density	Comp. str. test age	Compressive strength	Migration test age	CEM I	CEM II/A-D	CEM II/A-S	CEM II/B-S	CEM II/B-V	D_{nssm}
count	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91
mean	0.32	301.06	70.07	8.86	1.24	775.47	924.47	0.12	2200.61	20.15	40.22	27.01	0.53	0.02	0.29	0.05	0.11	8.16
std	0.06	91.39	64.28	23.92	5.4	124.66	271.54	0.25	182.08	10.22	9.59	4.78	0.5	0.15	0.45	0.23	0.31	2.98
min	0.22	186.88	0	0	0	517	372.82	0	1620	7	21.79	3	0	0	0	0	0	4.75
25 %	0.28	234.64	0	0	0	701.29	913.94	0	2230.89	7	32.71	28	0	0	0	0	0	6.25
50 %	0.32	261.04	83.65	0	0	753.46	1056.03	0	2274.62	28	40.38	28	1	0	0	0	0	7.54
75 %	0.36	363.38	121.62	0	0	809.53	1089.26	0	2303.45	28	46.5	28	1	0	1	0	0	9.2
max	0.45	500	186.88	100	29.66	1043	1187.15	0.7	2402.77	28	64.6	29	1	1	1	1	1	19.42

Table 6
Hyperparameters that delivers the best results.

	Hyperparameters				
Model	n_estimators	max_depth	learning_rate	booster	gamma
Model I	300	2	0.3	gbtree	0.0001
Model II	100	10	0.1	gbtree	0.01
Model III	100	2	0.1	gbtree	0.01
Model IV	100	4	0.3	gbtree	0.01

distributions of the four models, the median (middle quantile, 50th percentile) is very close to zero, confirming the models are performing rationally well.

The performance of all models on unseen data is also evaluated using four statistical measures, namely MSE, RMSE, MAE, and R^2, and the results are presented in Table 7. The smaller the statistical errors of MSE, RMSE, and MAE, the better the model performs. However, in the case of R^2, the situation is reversed. From Table 7 it can be seen that Model IV outperforms the rest of the models with validation errors of (MSE = 0.307, RMSE = 0.554, MAE = 0.416, and R^2 = 0.963). The superiority of this model corroborated that the fresh and hardened concrete properties (fresh density and compressive strength examined at early age) considered in the models are meaningful in predicting the chloride migration coefficient. This is because these properties are affected by factors other than the mix ingredients, such as compaction, which has a major impact on the strength, density, and permeability of concrete. Indeed, the features describing only the constitutes of the concrete mix are also sufficient to predict D_{nssm} with some accuracy, as demonstrated by the performance of Model I. The finding of fresh and hardened concrete properties examined at early age as influential features is supported by previous research [63]. The authors concluded that tests on fresh and hardened concrete, conducted at an early age, were effective in predicting chloride penetration into concrete.

The performance of the XGBoost is also compared to commonly used tree-based regression algorithms, which are decision tree and ensemble models (Random Forest, AdaBoost, Gradient Boosting, and Bagging). The statistical performance indicators (MSE, MAE, RMSE, and R^2) of all the algorithms are shown in Fig. 10. It can be seen that Model IV has the smallest MSE, MAE, and RMSE errors and the highest R-squared score of all the algorithms. This implies that Model IV is the best performing model of all algorithms, confirming that the fresh and hardened properties are the powerful predictors of D_{nssm}. Model II is the second closest best model. These two top performing models are obtained by using the XGBoost algorithm. Model III and Model I are the third and fourth performance models, respectively. It is recalled that in addition to the mix ingredients, Models II and III include additional input features describing the fresh density and comprehensive strength properties of the concrete, respectively. Though Model II outperforms Model III, it cannot be concluded that fresh density is a better predictor than comprehensive strength because the instances of the models' remaining input features are not identical. The random forest algorithm works best in the case of Model I and bagging in the case of Model III. The RMSEs of all the algorithms in the best performing model, Model IV, are shown in Fig. 11. XGBoost's RMSE is 1.73 and 3.3 times lower than the second best (gradient boosting) and worst (bagging) algorithms, respectively. All of this show that the XGBoost has a balanced trade-off between bias and variance errors, corroborating its generalization abilities.

The high generalization ability of the developed models, especially

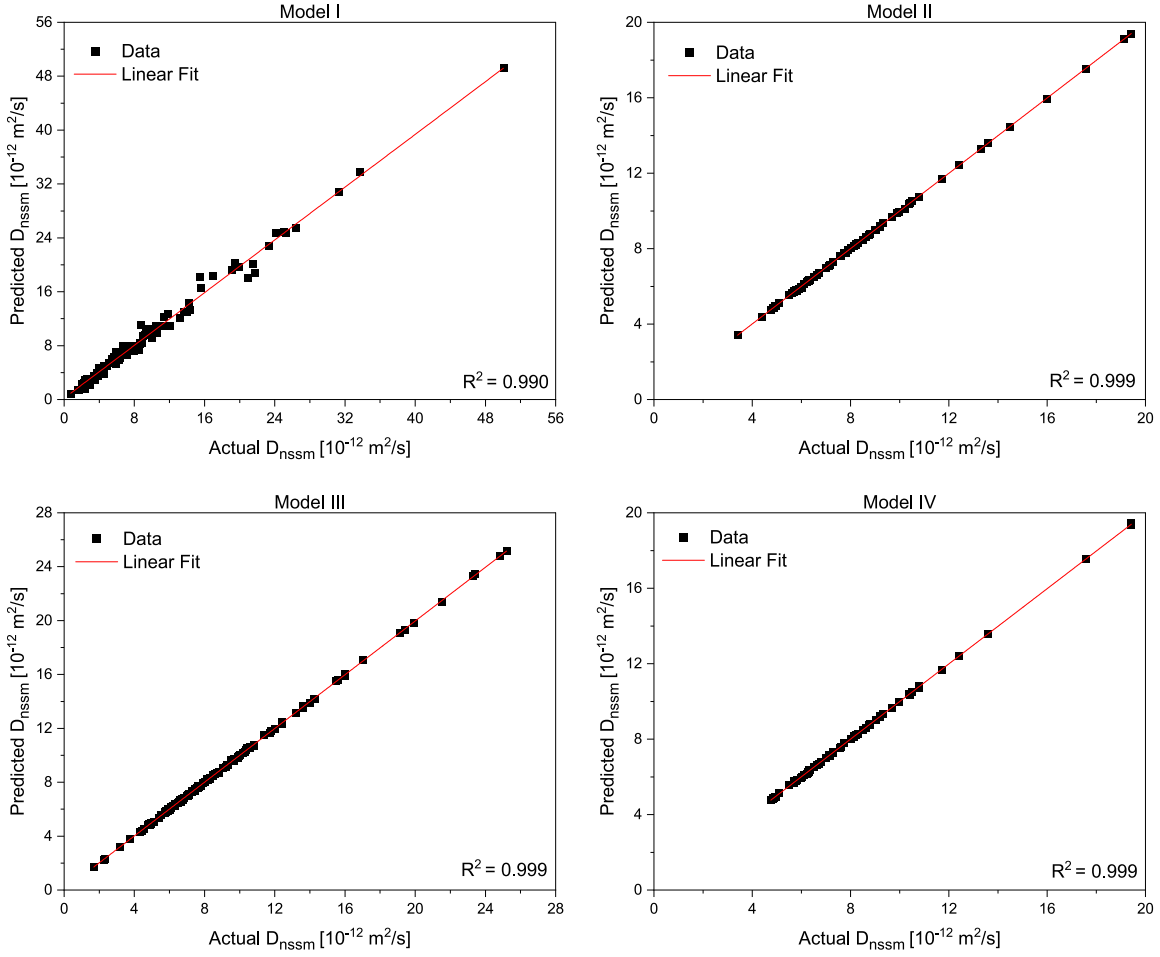

Fig. 6. Regression plots with the validation R-square scores of all models during the training phase.

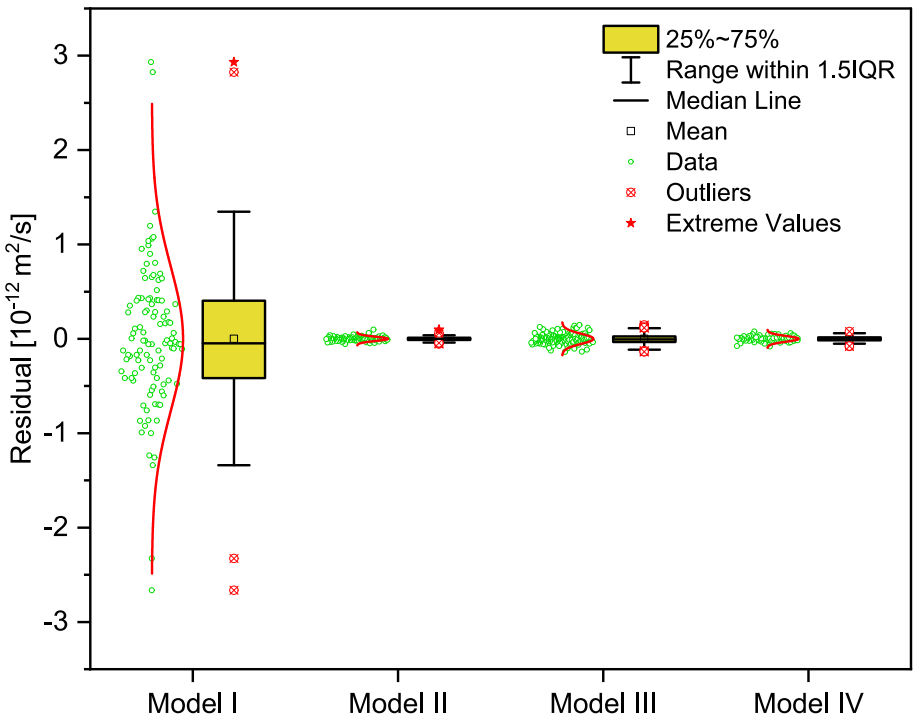

Fig. 7. Boxplots of the training residuals with the data points and distribution curves.

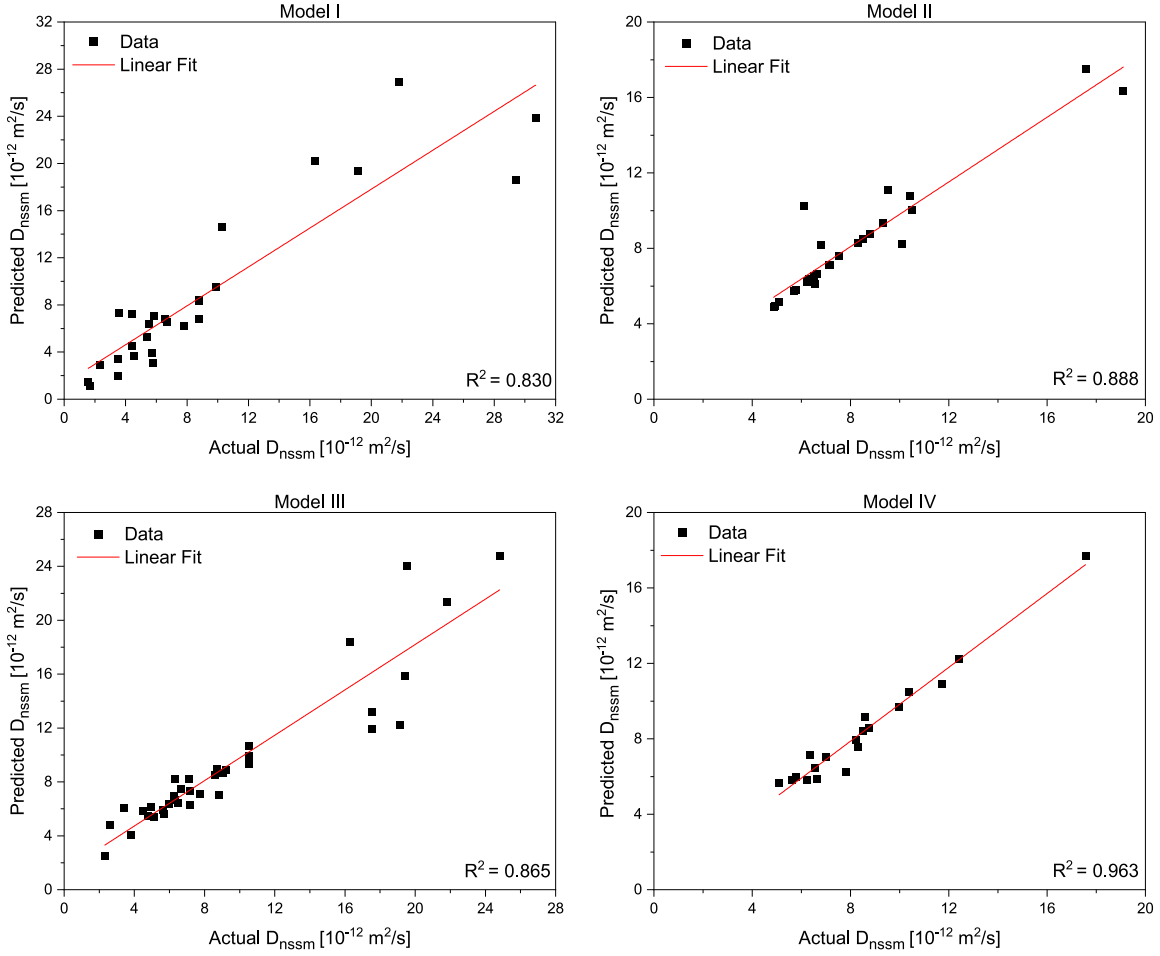

Fig. 8. Actual and predicted D_{nssm}.

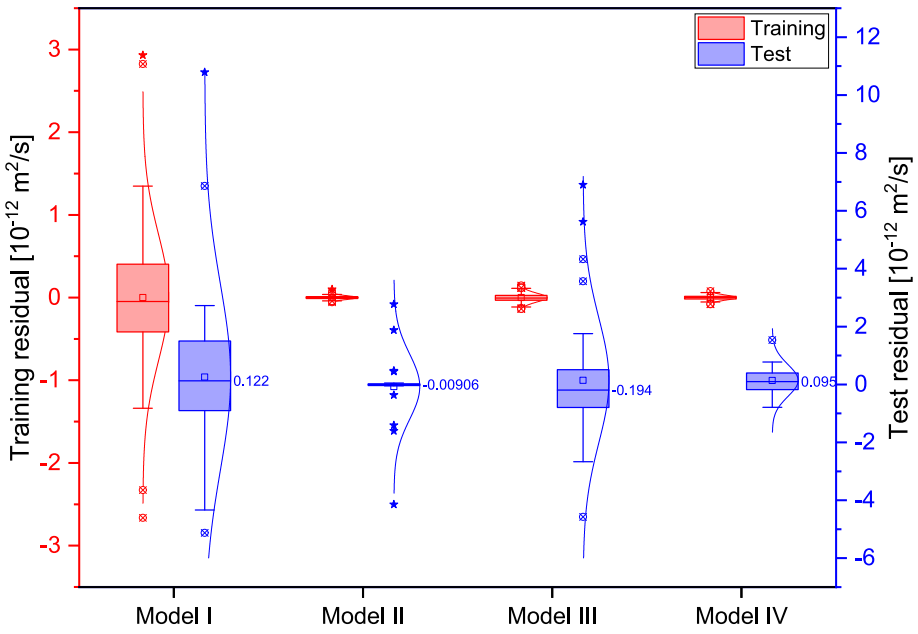

Fig. 9. Boxplots of the training and test residuals with distribution curves.

Table 7
Statistical validation metrics of the four models.

Models	MSE	RMSE	MAE	R^2
Model I	9.992	3.161	1.982	0.830
Model II	1.243	1.115	0.500	0.888
Model III	4.576	2.139	1.350	0.865
Model IV	0.307	0.554	0.416	0.963

Model IV, confirms their applicability in predicting the non-steady-state migration coefficient of concrete. Because the models were developed using a variety of concrete mixes from around the world, concrete designers can use them to evaluate the performance of their designed concrete against chloride resistance. It also has an economic implication as it helps design optimal concrete mixes without the need for advanced laboratory testing that is labor and resource intensive. In fact, the same approach used in this work could be used to develop reliable, universal and robust models to predict the chloride diffusivity of concrete, performed by different procedures and/or other important, time-consuming and resource-intensive tests.

The performance of all tested algorithms in this work varies from one model to another, as shown in Fig. 12. Examining the ability of different machine learning algorithms to predict the D_{nssm} is crucial to identifying other algorithms that might perform the best. This is because the relative predictive power of any machine learning algorithm is primarily determined by the specifics of the problems being considered. It's impossible to identify the powerful algorithms that excel at a given problem without experimenting. Using more representative data could also help further improve model performance. In fact, obtaining information from published studies was not an easy task. The use of multiple units of measurement, the lack of certain features, and the different methods of providing information were all common problems. This is because individual research conducted in different countries could only account for a limited number of features. Translating all of the data into appropriate units and formats to produce well-rounded data was time-consuming and required much attention. To address such a problem an open data exchanging platform that allows academia and/or the concrete community to share data in a sort of a standard format is required. In fact, with the advent of sensor technology, sensors to monitor the durability of concrete structures could increase in the years to come, leading to an influx of data [64,65]. Therefore, for the concrete industry to reap the significant benefits of machine learning, open data in a machine-readable format that can be freely shared by domain experts is required. With this, the concrete/ construction industry will no longer be an outlier in the ubiquitous digital revolution.

7. Conclusions

This study developed four machine learning-based chloride migration coefficient prediction models using the XGBoost algorithm. The primary distinction between the four models is the type of input features used. The model I makes use of input features that describe the concrete mix ingredients. Model II includes the same features as Model I plus a fresh density. Model III included the same features as Model I plus two features (strength test age and compressive strength). Model IV contains all of the feature types found in the previous three models. A large database of experimental data investigating the non-steady-state migration coefficient for a variety of concrete types was created using data gathered from research projects and previously published studies. All models were validated using previously unseen data. The statistical performance indicators MSE, RMSE, MAE, and R^2, were utilized to assess the accuracy of the models, and the results confirmed that all XGBoost-based models were able to predict D_{nssm} with a rationally small error. The performance of the XGBoost models was also compared to other tree-based algorithms. It works best on Model II and IV. The superiority of Model IV corroborates that the properties of fresh and hardened concrete, determined at an early age, are influential predictors of D_{nssm}. The features describing only the concrete mix components are also sufficient to predict the chloride migration coefficient with reasonable accuracy. The model is of great practical importance. Because the model was developed using a variety of concrete mixes from around the globe, it can be used by concrete designers from all over the world to evaluate the performance of their designed concrete against chloride resistance. It also has economic implications as it helps design optimal concrete mixes without the need for labor-intensive and resource-consuming advanced laboratory testing. This allows the models to be used to replace laboratory testing of the chloride migration coefficient of concrete, saving costs, time, and resources.

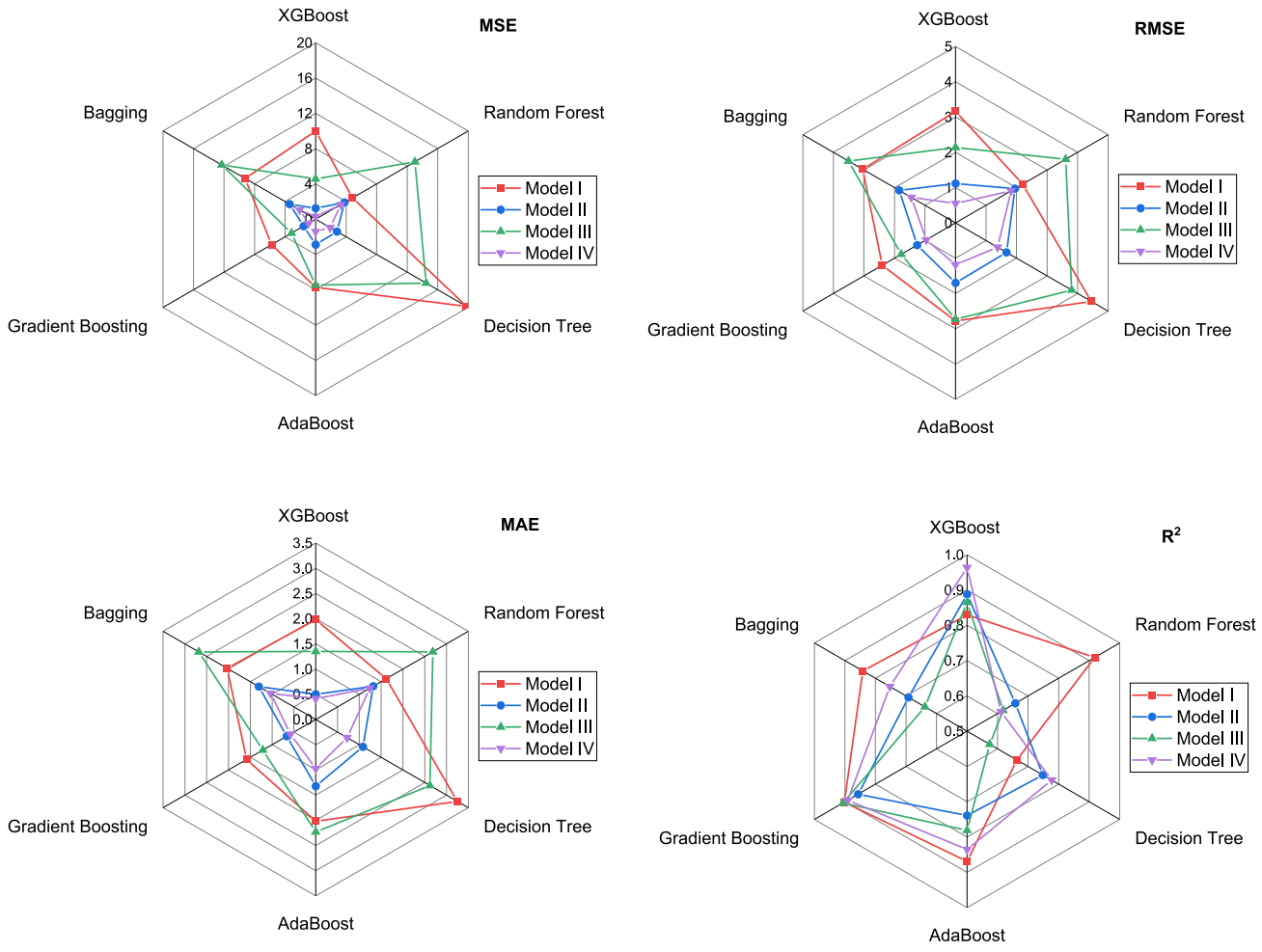

Fig. 10. Performance comparison of XGBoost with other tree-based algorithms.

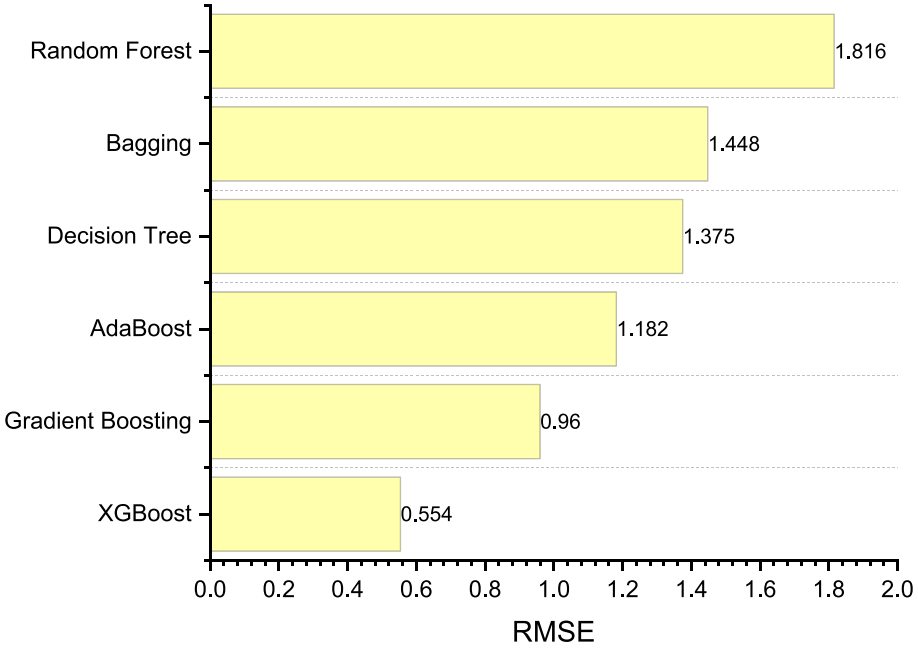

Fig. 11. Root-mean-square errors of all algorithms used in the case of Model IV.

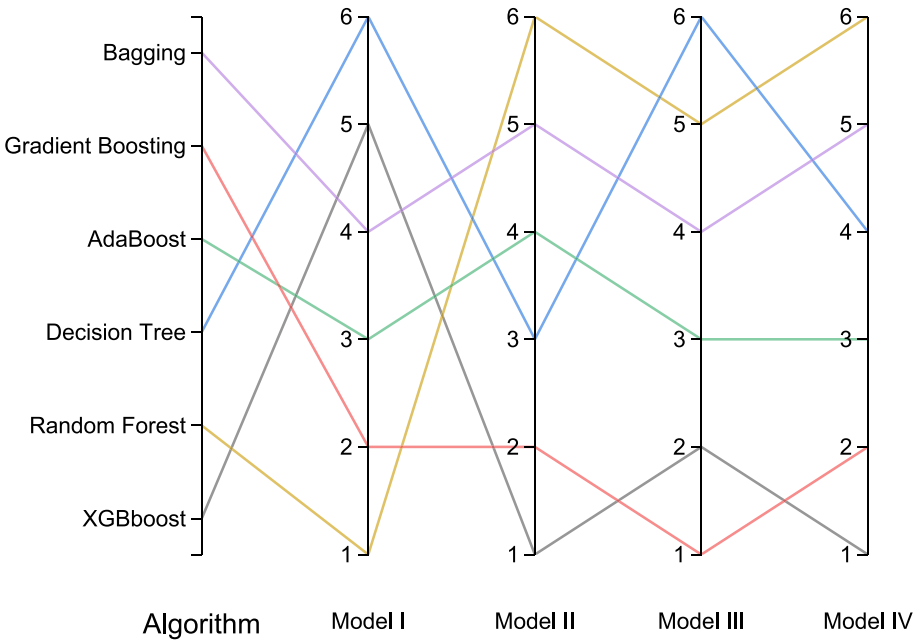

Fig. 12. Performance ranking of the algorithms in each model (1 = top and 6 = low performers).

CRediT authorship contribution statement

Woubishet Zewdu Taffese: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Leonardo Espinosa-Leal:** Writing – review & editing, Project administration, Funding acquisition.

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.

Data availability

The data used in this work are included as a supplementary file.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conbuildmat.2022.128566.

References

[1] J. Pontes, J.A. Bogas, S. Real, A. Silva, The rapid chloride migration test in assessing the chloride penetration resistance of normal and lightweight concrete, Appl. Sci. 11 (2021) 7251, https://doi.org/10.3390/app11167251.
[2] L. Tang, L.-O. Nilsson, P.A.M. Basheer, Resistance of concrete to chloride ingress: Testing and modelling, CRC Press, Boca Raton, FL (2012), https://doi.org/10.1201/b12603.
[3] T.S. Nguyen, S. Lorente, M. Carcasses, Effect of the environment temperature on the chloride diffusion through CEM-I and CEM-V mortars: An experimental study, Constr. Build. Mater. 23 (2009) 795–803, https://doi.org/10.1016/j.conbuildmat.2008.03.004.
[4] H. Ye, X. Jin, C. Fu, N. Jin, Y. Xu, T. Huang, Chloride penetration in concrete exposed to cyclic drying-wetting and carbonation, Constr. Build. Mater. 112 (2016) 457–463, https://doi.org/10.1016/j.conbuildmat.2016.02.194.
[5] X. Zhu, G. Zi, Z. Cao, X. Cheng, Combined effect of carbonation and chloride ingress in concrete, Constr. Build. Mater. 110 (2016) 369–380, https://doi.org/10.1016/j.conbuildmat.2016.02.034.
[6] M. Torres-Luque, E. Bastidas-Arteaga, F. Schoefs, M. Sánchez-Silva, J.F. Osma, Non-destructive methods for measuring chloride ingress into concrete: State-of-the-art and future challenges, Constr. Build. Mater. 68 (2014) 68–81, https://doi.org/10.1016/j.conbuildmat.2014.06.009.
[7] Y.C. Choi, B. Park, G.S. Pang, K.M. Lee, S. Choi, Modelling of chloride diffusivity in concrete considering effect of aggregates, Constr. Build. Mater. 136 (2017) 81–87, https://doi.org/10.1016/j.conbuildmat.2017.01.041.
[8] V. Elfmarkova, P. Spiesz, H.J.H. Brouwers, Determination of the chloride diffusion coefficient in blended cement mortars, Cem. Concr. Res. 78 (2015) 190–199, https://doi.org/10.1016/j.cemconres.2015.06.014.
[9] K. Audenaert, Q. Yuan, G. De Schutter, On the time dependency of the chloride migration coefficient in concrete, Constr. Build. Mater. 24 (2010) 396–402, https://doi.org/10.1016/j.conbuildmat.2009.07.003.
[10] ASTM C1556 - 11a, Standard test method for determining the apparent chloride diffusion coefficient of cementitious mixtures by bulk diffusion, ASTM, West Conshohocken, PA, 2016.
[11] N.T. Build, 443, Concrete, hardened: Accelerated chloride penetration, NORDTEST (2010).
[12] N.T. Build, 492, Concrete, mortar and cement-based repair materials: Chloride migration coefficient from non-steady-state migration experiments, NORDTEST (1999).
[13] S.E. Chidiac, M. Shafikhani, Phenomenological model for quantifying concrete chloride diffusion coefficient, Constr. Build. Mater. 224 (2019) 773–784, https://doi.org/10.1016/j.conbuildmat.2019.07.006.
[14] K.A. Riding, M.D.A. Thomas, K.J. Folliard, Apparent diffusivity model for concrete containing supplementary cementitious materials, ACI Mater. J. 110 (2013) 705–713, https://doi.org/10.14359/51686338.
[15] J.A. Bogas, A. Gomes, Non-steady-state accelerated chloride penetration resistance of structural lightweight aggregate concrete, Cem. Concr. Compos. 60 (2015) 111–122, https://doi.org/10.1016/j.cemconcomp.2015.04.001.
[16] G. Sun, W. Sun, Y. Zhang, Z. Liu, Multi-scale modeling of the effective chloride ion diffusion coefficient in cement-based composite materials, J. Wuhan Univ. Technol. Mater. Sci. Ed. 27 (2012) 364–373, https://doi.org/10.1007/s11595-012-0467-6.
[17] M. Shafikhani, S.E. Chidiac, Quantification of concrete chloride diffusion coefficient – A critical review, Cem. Concr. Compos. 99 (2019) 225–250, https://doi.org/10.1016/j.cemconcomp.2019.03.011.
[18] W.Z. Taffese, K.A. Abegaz, Artificial intelligence for prediction of physical and mechanical properties of stabilized soil for affordable housing, Appl. Sci. 11 (2021) 7503, https://doi.org/10.3390/app11167503.
[19] M. Saadat, M. Bayat, Prediction of the unconfined compressive strength of stabilised soil by Adaptive Neuro Fuzzy Inference System (ANFIS) and Non-Linear Regression (NLR), Geomech. Geoengin. 17 (2022) 80–91, https://doi.org/10.1080/17486025.2019.1699668.
[20] W.Z. Taffese, K.A. Abegaz, Prediction of compaction and strength properties of amended soil using machine learning, Buildings. 12 (2022) 613, https://doi.org/10.3390/buildings12050613.
[21] S. Pan, Z. Zheng, Z. Guo, H. Luo, An optimized XGBoost method for predicting reservoir porosity using petrophysical logs, J. Pet. Sci. Eng. 208 (2022), 109520, https://doi.org/10.1016/j.petrol.2021.109520.
[22] W.Z. Taffese, E. Sistonen, J. Puttonen, Prediction of concrete carbonation depth using decision trees, in: 23rd Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn., i6doc.com publisher, 2015.
[23] W. Dong, Y. Huang, B. Lehane, G. Ma, XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring, Autom. Constr. 114 (2020), 103155, https://doi.org/10.1016/j.autcon.2020.103155.

[24] W.Z. Taffese, E. Sistonen, J. Puttonen, CaPrM: Carbonation prediction model for reinforced concrete using machine learning methods, Constr. Build. Mater. 100 (2015) 70–82, https://doi.org/10.1016/j.conbuildmat.2015.09.058.
[25] A. Lavercombe, X. Huang, S. Kaewunruen, Machine learning application to eco-friendly concrete design for decarbonisation, Sustainability. 13 (2021) 13663, https://doi.org/10.3390/su132413663.
[26] 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.
[27] R. Cai, T. Han, W. Liao, J. Huang, D. Li, A. Kumar, H. Ma, Prediction of surface chloride concentration of marine concrete using ensemble machine learning, Cem. Concr. Res. 136 (2020), 106164, https://doi.org/10.1016/j.cemconres.2020.106164.
[28] A. Ahmad, F. Farooq, K.A. Ostrowski, K. Śliwa-Wieczorek, S. Czarnecki, Application of novel machine learning techniques for predicting the surface chloride concentration in concrete containing waste material, Materials (Basel). 14 (2021) 2297, https://doi.org/10.3390/ma14092297.
[29] O.A. Mohamed, M. Ati, W. Al Hawat, Implementation of artificial neural networks for prediction of chloride penetration in concrete, Int. J. Eng. Technol. 7 (2018) 47–52, https://doi.org/10.14419/ijet.v7i2.28.12880.
[30] M. Najimi, N. Ghafoori, M. Nikoo, Modeling chloride penetration in self-consolidating concrete using artificial neural network combined with artificial bee colony algorithm, J. Build. Eng. 22 (2019) 216–226, https://doi.org/10.1016/j.jobe.2018.12.013.
[31] S. Kumar, B. Rai, R. Biswas, P. Samui, D. Kim, Prediction of rapid chloride permeability of self-compacting concrete using Multivariate Adaptive Regression Spline and Minimax Probability Machine Regression, J. Build. Eng. 32 (2020), 101490, https://doi.org/10.1016/j.jobe.2020.101490.
[32] N.-D. Hoang, C.-T. Chen, K.-W. Liao, Prediction of chloride diffusion in cement mortar using Multi-Gene Genetic Programming and Multivariate Adaptive Regression Splines, Measurement 112 (2017) 141–149, https://doi.org/10.1016/j.measurement.2017.08.031.
[33] O.A. Hodhod, H.I. Ahmed, Developing an artificial neural network model to evaluate chloride diffusivity in high performance concrete, HBRC J. 9 (2013) 15–21, https://doi.org/10.1016/j.hbrcj.2013.04.001.
[34] L. Yao, L. Ren, G. Gong, Evaluation of chloride diffusion in concrete using PSO-BP and BP neural network, IOP Conf. Ser. Earth Environ. Sci. 687 (2021), 012037, https://doi.org/10.1088/1755-1315/687/1/012037.
[35] J.M.P.Q. Delgado, F.A.N. Silva, A.C. Azevedo, D.F. Silva, R.L.B. Campello, R. L. Santos, Artificial neural networks to assess the useful life of reinforced concrete elements deteriorated by accelerated chloride tests, J. Build. Eng. 31 (2020), 101445, https://doi.org/10.1016/j.jobe.2020.101445.
[36] M. Marks, D. Jóźwiak-NiedzWiedzka, M.A. Glinicki, Automatic categorization of chloride migration into concrete modified with CFBC ash, Comput. Concr. 9 (2012) 375–387, https://doi.org/10.12989/cac.2012.9.5.375.
[37] M. Marks, M.A. Glinicki, K. Gibas, Prediction of the chloride resistance of concrete modified with high calcium fly ash using machine learning, Materials (Basel). 8 (2015) 8714–8727, https://doi.org/10.3390/ma8125483.
[38] B. Quinto, Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More, Apress (2020), https://doi.org/10.1007/978-1-4842-5669-5.
[39] J.G. Xu, S.Z. Chen, W.J. Xu, Z. Sen Shen, Concrete-to-concrete interface shear strength prediction based on explainable extreme gradient boosting approach, Constr. Build. Mater. 308 (2021), 125088, https://doi.org/10.1016/j.conbuildmat.2021.125088.
[40] A. Shehadeh, O. Alshboul, R.E. Al Mamlook, 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), 103827, https://doi.org/10.1016/j.autcon.2021.103827.
[41] L. Tang, H.E. Sørensen, Precision of the Nordic test methods for measuring the chloride diffusion/migration coefficients of concrete, Mater. Struct. Constr. 34 (2001) 479–485, https://doi.org/10.1007/bf02486496.
[42] M.T. Hasholt, O.M. Jensen, Chloride migration in concrete with superabsorbent polymers, Cem. Concr. Compos. 55 (2015) 290–297, https://doi.org/10.1016/j.cemconcomp.2014.09.023.
[43] H. Kuosa, Concrete durability field testing in DuraInt-project: Field and laboratory results 2007 - 2010, Espoo, 2011.
[44] F.K. Sell Junior, G.B. Wally, F.R. Teixeira, F.C. Magalhães, Experimental assessment of accelerated test methods for determining chloride diffusion coefficient in concrete, Rev. IBRACON Estruturas e Mater. 14 (2021), https://doi.org/10.1590/s1983-41952021000400007.
[45] H.B. Hou, G.Z. Zhang, Assessment on chloride contaminated resistance of concrete with non-steady-state migration method, J. Wuhan Univ. Technol. Mater. Sci. Ed. 19 (2004) 6, https://doi.org/10.1007/bf02841355.
[46] R.W. Shiu, C.C. Yang, Evaluation of migration characteristics of opc and slag concrete from the rapid chloride migration test, J. Mar. Sci. Technol. 28 (2020) 69–79, https://doi.org/10.6119/JMST.202004_28(2).0001.
[47] M. Maes, E. Gruyaert, N. De Belie, Resistance of concrete with blast-furnace slag against chlorides, investigated by comparing chloride profiles after migration and diffusion, Mater. Struct. 46 (2013) 89–103, https://doi.org/10.1617/s11527-012-9885-3.
[48] J. Jain, N. Neithalath, Electrical impedance analysis based quantification of microstructural changes in concretes due to non-steady state chloride migration, Mater. Chem. Phys. 129 (2011) 569–579, https://doi.org/10.1016/j.matchemphys.2011.04.057.
[49] X. Liu, K.S. Chia, M.H. Zhang, Water absorption, permeability, and resistance to chloride-ion penetration of lightweight aggregate concrete, Constr. Build. Mater. 25 (2011) 335–343, https://doi.org/10.1016/j.conbuildmat.2010.06.020.
[50] S. Real, J.A. Bogas, J. Pontes, Chloride migration in structural lightweight aggregate concrete produced with different binders, Constr. Build. Mater. 98 (2015) 425–436, https://doi.org/10.1016/j.conbuildmat.2015.08.080.
[51] C. Naito, J. Fox, P. Bocchini, M. Khazaali, Chloride migration characteristics and reliability of reinforced concrete highway structures in Pennsylvania, Constr. Build. Mater. 231 (2020), 117045, https://doi.org/10.1016/j.conbuildmat.2019.117045.
[52] J.-I. Park, K.-M. Lee, S.-O. Kwon, S.-H. Bae, S.-H. Jung, S.-W. Yoo, Diffusion Decay Coefficient for Chloride Ions of Concrete Containing Mineral Admixtures, Adv. Mater. Sci. Eng. 11 (2016 (2016)) pages, https://doi.org/10.1155/2016/2042918.
[53] X. Liu, H. Du, M.H. Zhang, A model to estimate the durability performance of both normal and light-weight concrete, Constr. Build. Mater. 80 (2015) 255–261, https://doi.org/10.1016/j.conbuildmat.2014.11.033.
[54] R. Van Noort, M. Hunger, P. Spiesz, Long-term chloride migration coefficient in slag cement-based concrete and resistivity as an alternative test method, Constr. Build. Mater. 115 (2016) 746–759, https://doi.org/10.1016/j.conbuildmat.2016.04.054.
[55] R.M. Ferreira, J.P. Castro-Gomes, P. Costa, R. Malheiro, Effect of metakaolin on the chloride ingress properties of concrete, KSCE J. Civ. Eng. 20 (2016) 1375–1384, https://doi.org/10.1007/s12205-015-0131-8.
[56] A. Pilvar, A.A. Ramezanianpour, H. Rajaie, S.M.M. Karein, Practical evaluation of rapid tests for assessing the Chloride resistance of concretes containing Silica Fume, Comput. Concr. 18 (2016) 793–806, https://doi.org/10.12989/cac.2016.18.6.793.
[57] J. Liu, X. Wang, Q. Qiu, G. Ou, F. Xing, Understanding the effect of curing age on the chloride resistance of fly ash blended concrete by rapid chloride migration test, Mater. Chem. Phys. 196 (2017) 315–323, https://doi.org/10.1016/j.matchemphys.2017.05.011.
[58] En,, 197–1, Cement- Part 1: Composition, specifications and conformity criteria for common cements, CEN (2011).
[59] K.M. Sunderland, D. Beaton, J. Fraser, D. Kwan, P.M. McLaughlin, M. Montero-Odasso, A.J. Peltsch, F. Pieruccini-Faria, D.J. Sahlas, R.H. Swartz, R. Bartha, S. E. Black, M. Borrie, D. Corbett, E. Finger, M. Freedman, B. Greenberg, D.A. Grimes, R.A. Hegele, C. Hudson, A.E. Lang, M. Masellis, W.E. McIlroy, D.G. Munoz, D. P. Munoz, J.B. Orange, M.J. Strong, S. Symons, M.C. Tartaglia, A. Troyer, L. Zinman, S.C. Strother, M.A. Binns, The utility of multivariate outlier detection techniques for data quality evaluation in large studies: An application within the ONDRI project, BMC Med. Res. Method. 19 (2019) 1–16, https://doi.org/10.1186/s12874-019-0737-5.
[60] K. Varmuza, P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics, CRC Press, 2016.
[61] U. Stańczyk, Feature evaluation by filter, wrapper, and embedded approaches, in: U. Stańczyk, L.C. Jain (Eds.), Featur. Sel. Data Pattern Recognit, Springer-Verlag, Berlin, 2015, pp. 29–44, https://doi.org/10.1007/978-3-662-45620-0.
[62] S. Marsland, Machine learning: an algorithmic perspective, Chapman and Hall/CRC, Boca Raton, FL, 2011.
[63] 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.
[64] W.Z. Taffese, E. Nigussie, J. Isoaho, Internet of things based durability monitoring and assessment of reinforced concrete structures, Procedia Comput. Sci. 155 (2019) 672–679, https://doi.org/10.1016/j.procs.2019.08.096.
[65] W.Z. Taffese, E. Nigussie, Autonomous corrosion assessment of reinforced concrete structures: Feasibility study, Sensors (Switzerland). 20 (2020) 6825, https://doi.org/10.3390/s20236825.