ARCADA

THIS IS A SELF-ARCHIVED VERSION OF THE ORIGINAL PUBLICATION

The self-archived version is a publisher‘s pdf of the original publication. NB. The self-archived version may differ from the original in pagination, typographical details and illustrations.

To cite this, use the original publication:

Taffese, W. Z., & Espinosa-Leal, L. (2022). Prediction of chloride resistance level of concrete using machine learning for durability and service life assessment of building structures. Journal of Building Engineering, 60, 24 p.

DOI: https://doi.org/10.1016/j.jobe.2022.105146

Permanent link to the self-archived copy:

All material supplied via Arcada’s self-archived publications collection in Theseus repository is protected by copyright laws. Use of all or part of any of the repository collections is permitted only for personal non-commercial, research or educational purposes in digital and print form. You must obtain permission for any other use.

JAN-MAGNUS JANSSONS PLATS 1, FIN-00550 HELSINGFORS, TEL: +358 (0)20 769 9699 FAX: +358 (0)20 769 9622 www.arcada.fi

Journal of Building Engineering 60 (2022) 105146

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Prediction of chloride resistance level of concrete using machine learning for durability and service life assessment of building structures

Woubishet Zewdu Taffese *, Leonardo Espinosa-Leal

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

ARTICLE INFO

Keywords:
Coastal buildings
Chloride diffusion
Chloride resistance
Non-steady-migration coefficients
Machine learning
Classification
Prediction
Service life
Durability

ABSTRACT

The resistance of concrete to chloride penetration determines the durability and service life of reinforced concrete building structures in coastal or chloride-laden environments. This work adopted five machine learning algorithms, naïve bayes, k-nearest neighbors, decision trees, support vector machine, and random forests, to predict the chloride resistance level of concrete based on its ingredients, considering two scenarios. The first scenario considers all features describing the mix components, whereas the second scenario considers only a subset of the features. All models are validated by performing intensive evaluation matrices using unseen data. The validation results confirm that the developed models predict the level of chloride resistance of concrete with high accuracy. Of all the algorithms, the support vector machine performed best, with 89% and 88% accuracy in the first and second scenarios, respectively.

1. Introduction

Concrete is a popular building material in the coastal region, as evidenced by the large number of coastal buildings, or structures on or near the coast. The development and utilization of coastal regions has increased significantly in recent decades, allowing for an increase in economic activity and employment opportunities, which leads to an increase in the human population along the coastal regions. According to the United Nations, approximately 40% of the world's population lived within 100 km of the coast in 2017 [1]. This figure is expected to rise as the trend of population growth and migration from inland to coastal areas continues, increasing the demand for residential and office buildings in the coastal region. Concrete will continue to be used in coastal construction because it is abundant, resilient, durable, and affordable, and it can be used in an infinite number of ways.

Although concrete is a durable building material, coastal buildings or even non-coastal buildings subjected to deicing salts, such as parking lots, can suffer from chloride attack. In coastal areas, marine aerosol, produced by the rupture of small bubbles on the sea surface and carried far inland by the wind, is the main source of chloride, which attacks the concrete structure [2]. Deicing salts, which are used to melt ice on roads during the winter, are a source of chloride that attacks concrete in non-coastal areas. According to Ref. [3], chloride from deicing salt was detected as high as the 60th floor of a reinforced concrete building structure located 1.9 km from a busy highway.

The penetration of chloride ions into the pores of concrete plays a crucial role in the physical and chemical processes associated with the deterioration of the microstructure of concrete and the corrosion of reinforcing bars in concrete. Despite the fact that concrete

* Corresponding author.
E-mail addresses: woubishet.taffese@arcada.fi (W.Z. Taffese), leonardo.espinosaleal@arcada.fi (L. Espinosa-Leal).

https://doi.org/10.1016/j.jobe.2022.105146
Received 1 June 2022; Received in revised form 15 July 2022; Accepted 16 August 2022
Available online 30 August 2022
2352-7102/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

is typically alkaline, with a pore solution pH of 12–13, it passivates the embedded rebars and prevents corrosion. Nevertheless, depassivation occurs when the chloride concentration at the reinforcement bar reaches a certain level, causing corrosion to initiate and ultimately resulting in a significant reduction in the serviceability, safety, and economy of the coastal structures [2,4]. Hence, understanding the chloride transport process is critical to extending the lifespan of structures in coastal environments. The transport of chloride ions in concrete, on the other hand, is a complex physicochemical process involving various transport mechanisms such as diffusion, capillary suction, and permeation [5–7]. This process is simplified by assuming that diffusion is the primary mechanism for chloride transport into the concrete medium. Fick's second law of diffusion, which assumes a constant diffusion coefficient, is a universally applicable mathematical model for determining non-steady state diffusion, as represented by Equation (1).

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

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

The chloride ion diffusion coefficient, D, is defined as the transfer rate of chloride ions across a unit area of a concrete section divided by the concentration gradient in space at the section. Because it is the primary durability indicator, several service-life prediction models, such as Life-365 [8] and DuraCrete [9], rely on its value to assess the service life of a specific coastal concrete building and to devise a cost-effective maintenance strategy. Concrete chloride diffusion coefficient is typically determined in a laboratory using various test procedures such as ASTM C1556–11 [10], NT Build 443 [11] and NT Build 492 [12]. Unlike the other tests, the Nordic standard NT Build 492 [12] test is completed in a relatively short period of time, making it a widely accepted test method. This test method is based on the migration of chloride ions into previously vacuum saturated concrete samples in anolyte solution (0.3 M NaOH) using an external electrical voltage (30 V DC) set between 10 and 60 V according to the electrical current in the scheme. The applied electrical potential causes chloride ions from a 10% NaCl solution (catholyte) to migrate into the concrete samples. After a certain amount of time, the samples are split axially and a silver nitrate solution (AgNO3) is sprayed onto the freshly fractured surface, where reacts with chloride ions on contact to form white insoluble silver chloride. This allows for the measurement of chloride penetration depths at 10 mm intervals across the cleaved surface, giving 5 to 7 valid depth readings [12,13]. As a representative value, the average of the chloride depths is used. After determining the chloride depth, the chloride migration coefficient is calculated using Equation (2) [12]. The diffusion coefficient determined by this method is called the "non-steady-state migration coefficient" or D_{nssm} (also in this study) to distinguish it from other diffusion tests.

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

where D_{nssm} is the non-steady-state migration coefficient, $\cdot 10^{-12}$ 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).

One of the key aspects of the durability design approach that extends the service life of the concrete structures exposed to a coastal or chloride-laden environment is to limit the movement of chloride ions into the concrete. This can be accomplished by designing the best combination of concrete mix ingredients capable of resisting chloride ion penetration for the intended service life and exposure conditions. In fact, designing the optimal mix ratio of sustainable concrete that meets the desired chloride resistance properties while using the least amount of cement is often challenging. This is mainly due to the rapid increase in the use of various supplementary cementitious materials (SCMs) (such as natural pozzolan, recycled ground glass pozzolan), and several industrial by-products (such as fly ash (FA), ground granulated blast furnace slag (GGBFS), and silica fume (SF)). In order to determine if the designed concrete achieves the desired level of chloride penetration resistance, a number of test concrete samples must be prepared and tested in a laboratory using specific test methods, usually NT Build 492. Even though this test method yields quick results, it is typically performed 28 days after the concrete samples are prepared. If the first attempt is unsuccessful, the entire process will be repeated, which will take more time and resources. As a result, experimentally determining the chloride resistance level of concrete for each project is difficult due to the high cost, time and resources involved. Therefore, a reliable, fast, uncomplicated, and cost-effective method to predict the level of the chloride penetration resistance of concrete with fewer processes is of paramount importance.

The development of a reliable and uncomplicated physically based method for predicting the chloride resistance level of concrete, which considers all influencing parameters is undoubtedly a complicated task, since the chloride permeability property of concrete is a function of numerous parameters that are mathematically complex to represent without considering several assumptions and simplifications. To counteract these issues, developing an artificial intelligence (AI)-driven system using state-of-the-art machine learning (ML) algorithms is a better alternative as they can solve complex problems involving many features without making assumptions.

The aim of this work is to provide a reliable method for ensuring the durability and service life of reinforced concrete buildings in coastal areas or those exposed to chloride-laden environments with minimal or even without the need for time-consuming and resource-intensive laboratory tests of chloride diffusion tests. The objectives are twofold: i) devising a machine learning based approach for predicting the degree of resistance of concrete to chloride penetration for the durability design and service-life assessment of chloride-exposed reinforced concrete buildings, and ii) investigating the importance of concrete mix components in predicting the resistance of concrete to chloride penetration and contributing to a better understanding of the durability and service life of reinforced concrete buildings in coastal region or subjected to chloride-loaded environments.

2. Related work

Obviously, a rapid and resource-efficient assessment of concrete's resistance to chloride penetration is critical to the durability of reinforced concrete structures exposed to the coastal environment and deicing salt. In this regard, there have been few studies that have used machine learning to predict the level of resistance to chloride penetration of concrete for specific types of concrete without the use of time-consuming and resource-intensive laboratory test procedures. For example, Marks et al. [14] adopted 20 machine learning algorithms to predict the degree of resistance to chloride penetration in concrete modified with high calcium fly ash (HCFA). The algorithms can be grouped into three types: i) Bayesian, ii) tree, and iii) rule. Cement content, fly ash with a high calcium content, water content, and specific surface of fly ash were used as predictors. The J48, a tree-based algorithm, performed best. Another relevant study conducted by Marks et al. [15] used the J48 algorithm to predict the chloride migration resistance level of concrete modified with circulating fluidized bed combustion (CFBC). They utilize cement content, fluidized fly ash obtained from two different sources, water content, air content, and compressive strength as predictors. According to the authors, the algorithm was suitable for classifying the degree of resistance of the concrete to chloride penetration. In addition to reporting the J48 algorithms' prediction ability, both studies [14,15] reported that the rules generated by the algorithm aid in evaluating the effect of the mix components in controlling the resistance of the concrete to chloride penetration.

Indeed, there have been studies that demonstrate the potential of machine learning algorithms to predict the diffusion coefficient of concrete [16–19] ultimately assisting in classifying the degree of resistance of the concrete to chloride penetration. For instance, artificial neural networks (ANNs) was used by Hodhod and Ahmed [16] to predict the chloride diffusivity of high-performance concrete (HPC). The four input parameters used in the model were cement content, water-to-binder (w/b) ratio, FA or GGBFS content, and age of cure. Based on the performance evaluation, the authors reported that the model predicted the chloride diffusion coefficient with high degree of accuracy. It also overcame the disadvantage of a lengthy laboratory testing period, making it a powerful, quick, and low-cost alternative for determining the chloride diffusivity of HPC. Yao et al. [17] describes the use of a backpropagation (BP) neural network and the particle swarm optimization (PSO) on the BP neural network to predict chloride penetration into concrete. The experimental concrete samples were produced using a variety of mineral admixtures (such as GGBS, FA, and SF). They use eight input layers for the network, representing the w/b ratio, the content of all binders, fine and coarse aggregates, and the age of cure. The authors claimed that the PSO-BP neural network outperforms the BP neural network in terms of prediction and is therefore an effective method for evaluating lifespan prediction for concrete structures. Hoang et al. [18] proposed multivariate adaptive regression splines (MARS) and multi-gene genetic programming (MGGP) to predict chloride diffusion in cement mortar. After comparing the performance of the models to that of ANN and least squares support vector regression (LSSVR), the authors concluded that the models outperformed them. Indeed, both MGGP and MARS can aid in the development of modelling equations with desirable prediction accuracy. In addition, they could help to the discovery of important parameters that control the diffusion of chloride ions in cement mortar. The authors also reported that the predictive equations generated by MARS and MGGP can aid decision makers during the design phase of concrete structures used in marine environments. In a study by Delgado et al. [19], ANN was used to predict the diffusion coefficient and chloride penetration of concrete samples under drying–wetting cycle conditions. Predictors include cement type, w/c, mineral additive type, age of cure, and number of drying-wetting cycles. According to the authors, the developed model maps the relationship between the considered predictor variables and thus accurately predicts both parameters.

3. Research significance

The attempted machine learning methods for predicting the level of resistance to chloride penetration and the chloride diffusion coefficient of concrete have yielded promising results. However, these methods have some limitations, which are listed as follows. i) all the models predict the chloride transport properties of certain types of concrete such as HPC, self-compacting concrete (SCC), normal weight concrete (NWC), and cement mortar. This is because the experimental data used to develop each model lacked a diverse range of concrete types. As a result, the models are not universally applicable. ii) most of previous works failed to consider all the parameters describing the properties of the ingredients. The transport of chlorides into the concrete pores is well known to be governed by the concrete microstructure, which is enormously complex and strongly influenced by the nature of the mix components and ratios. iii) only a few types of machine learning algorithms have been tested. Examining other types of machine learning algorithms is also important because the relative predictive power of any algorithm is determined primarily by the specifics of the problems being considered. Without experimentation, it is impossible to identify the powerful algorithms that might excel for a given problem.

The previously proposed models may fit the experimental data employed in their work, but they are unlikely to predict the level of concrete's resistance to chloride penetration or the chloride diffusion coefficient of concrete based on data from other experiment, resulting in inaccurate durability and service life assessments for reinforced concrete structures. All influencing parameters that govern the microstructure of various types of concrete must be taken into account in order to develop a reliable and universal model. Furthermore, many types of machine learning algorithms are to be experimented. To address the identified shortcomings, this study included a number of machine learning techniques that had not been previously used, as well as a large amount of data that included a variety of concrete types. In this study, significant optimization approaches are also used to improve the performance of the models that predict the level of chloride resistance of concrete. The subsequent section provides a detailed explanation of the algorithms used. The dataset and the applied optimization strategies are described in detail in Section 5.

4. Machine learning

Machine learning is a branch of artificial intelligence that involves the designing and implementing algorithms for recognizing

complicated patterns in data and make sensible decisions [20]. It has helped improve productivity across a wide of services and sectors. Although the use of machine learning in the building and infrastructure engineering sector is still in its infancy, its application has increased in recent years to solve several problems such as concrete durability [21–24], and geotechnics [25–27].

A machine learning method that maps an input to an output based on input-output pairs is called supervised learning. This type of learning can be divided into two categories based on the nature of the desired outcome (target feature): classification and regression. Classification is a type of supervised learning in which the target feature is represented by a finite set of discrete values. Regression, on the other hand, refers to situations where the value of the target feature is continuous. This work focuses on classification learning as it aims to predict the resistance class of concrete to chloride penetration. A machine learning classifier addresses the general problem of finding an approximation $\widehat{F}$ of an unknown function F defined from an input space Ω onto an unordered set of classes $\{w_1, \ldots, w_k\}$, given a training set: $T = \{(x^p, y^p = F(x^p)\}_{p=1}^{P} \subset \Omega \times \{w_1, \ldots, w_k\}$.

In this study, five commonly used classification algorithms that can handle multiclass classification tasks. These are: Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Trees, and Random Forests. The following sections explain their basic principle.

4.1. Naïve bayes

The naïve bayes (NB) algorithm is a simple multiclass linear classification algorithm based on Bayes' theorem. It is a classic example of how generative assumptions and parameter estimations can help to simplify the learning process. Consider the problem of predicting a label $y \in \{0, 1\}$ based on a vector of features $\mathbb{X} = (x_1, \ldots, x_d)$, where each x_i is in {0,1}. Recall that the Bayes optimal classifier is given by Equation (3) [28].

$$h_{Bayes}(\mathbb{X}) = argmax_{y \in \{0,1\}} P[Y = y | X = \mathbb{X}]. \tag{3}$$

To define the probability function $P[Y = y | X = \mathbb{X}]$ 2^d parameters are needed, each of which corresponds to $P[Y = 1 | X = \mathbb{X}]$ for a given value of $\mathbb{X} \in \{0, 1\}^d$. This means that the number of examples required grows exponentially with the number of features increases.

The NB approach naively assumes that each input feature is independent and ignores all possible correlations between features. That is, as expressed in Equation (4), changing the value of one feature does not directly affect the value of any of the other features used in the algorithm.

$$P[X = \mathbb{X} \,|Y = y] = \prod_{i=1}^{d} P[X_i = x_i \,|Y = y]. \tag{4}$$

With this assumption and the Bayes rule, the Bayes optimal classifier can be simplified even further, as shown in Equation (5), by significantly reducing the number of features to be learned.

$$\begin{aligned} h_{Bayes}(\mathbb{X}) &= argmax_{y \in \{0,1\}} P[Y = y | X = \mathbb{X}] \\ &= argmax_{y \in \{0,1\}} P[Y = y] P[X = \mathbb{X} | Y = y] \\ &= argmax_{y \in \{0,1\}} P[Y = y] \prod_{i=1}^{d} P[X_i = x_i \,|Y = y]. \end{aligned} \tag{5}$$

4.2. K-nearest neighbors

K-nearest neighbors (KNN) algorithms are among the most fundamental machine learning algorithms. The logic behind such a technique is based on the assumption that the features used to describe domain points are so relevant to their labels that neighboring points are likely to have the same label [28].

Let's assume that the instance domain, X, has a metric function ρ. In other words, a function $\rho : X \times X \rightarrow \mathbb{R}$ calculates the distance between any two X elements. For instance, if $X = \mathbb{R}^d$ then ρ can be the Euclidean distance, $\rho(\mathbb{X}, \mathbb{X}') = \mathbb{X} - \mathbb{X}' = \sqrt{\sum_{i=1}^{d} (x_i - x_i')^2}$.

Let $S = (\mathbb{X}_1, y_1), \ldots, (\mathbb{X}_m, y_m)$ represent a series of training instances. For each $\mathbb{X} \in X$, let $\pi_1(\mathbb{X}), \ldots, \pi_m(\mathbb{X})$ be a reordering of $\{1, \ldots, m\}$ based on their distance to $\mathbb{X}$, $\rho(\mathbb{X}, \mathbb{X}_i)$. That is, for all $i < m$,

$$\rho\left(\mathbb{X}, \mathbb{X}_{\pi_i(\mathbb{X})}\right) \leq \rho\left(\mathbb{X}, \mathbb{X}_{\pi_{i+1}(\mathbb{X})}\right). \tag{6}$$

The KNN rule for binary classification for a number k is defined as: input of a training sample $S = (\mathbb{X}_1, y_1), \ldots, (\mathbb{X}_m, y_m)$, and output for every point $\mathbb{X} \in X$, return the majority label among $\{y_{\pi_i(\mathbb{X})} : i \leq k\}$. When $k = 1$, the 1-NN rule is $h_s(X) = y_{\pi_1}(\mathrm{x})$.

4.3. Support vector machine

The support vector machine (SVM) works by determining the optimal hyperplane that maximizes the margin between two classes by dividing the data into points into separate classes as far apart as possible. Fig. 1 shows a representation of SVM obtaining the decision boundary (a line separating classes) for the classification of two separable groups. In two dimensions, the decision boundary can be defined by all pairs of points (x_1, x_2) for which $b_1 x_1 + x_1 b_2$ is a constant, with fixed coefficients b_1 and b_2. This can be written as

$b_0 + b_1x_1 + x_1b_2 = 0$, where the coefficient b_0 is the negative constant. In fact, based on the coefficient values, there could be an infinite number of possible lines.

At higher dimensions, the condition expressed in Equation (7) defines a hyperplane (decision plane) [29].

$$b_0 + b^T x = 0 \tag{7}$$

Suppose we have a training data set of n objects in the r-dimensional space, i.e., the vectors (points) $x_1, \ldots, x_n$. In the case of two-classes, we have the class membership information for each object, i.e., given by values y_i which are, for instance, either -1 (first class) or $+1$ (second class), for $i = 1, \ldots, n$. If the two classes are linearly separable, the hyperplane expressed in Equation (8) gives a perfect group separation:

$$y_i\left(b_0 + b^T x_i\right) > 0 \tag{8}$$

The data points defining the position of the separating hyperplanes can be viewed as vectors in the transformed space and are called support vectors. The resulting largest margin is labled by M in Fig. 1, and is represented by the two dashed lines. The solid line (hyperplane) represents the separating line located at a distance of $M/2$ between the dashed lines. When the distance between the groups is at a maximum, the hyperplane is assumed to be in its optimal position. The optimization problem can be formulated as shown in Equation (9)

$M \rightarrow$ max for coefficients b_0 and b with $b^T b = 1$

$$\text{subject to } y_i\left(b_0 + b^T x_i\right) \geq M/2 \text{ for } i = 1, \ldots, n \tag{9}$$

4.4. Decision trees

Decision tree (DT) is basically an acyclic connected graph structure comprises nodes, branches, and leaves as shown in Fig. 2. Each internal node of the decision tree represents a condition, and the predictive model chooses an appropriate branch to go the next node based on the current node splitting criterion. The model repeats this process until it reaches the leaf node, which contains the classification results (target class label). DT algorithms can use various splitting criteria, such as entropy, Gini-index, Chi-square test, and F-test. The entropy and the Gini-index are the commonly used criteria for categorical target. Entropy is a measure of randomness or uncertainty. The entropy of each split ranges from 0 to 1 and the lower the entropy, the less uniform the distribution, the purer the node. The entropy can be calculated using the formula given in Equation (10). The information gain, i.e., information that can increase the certainty level after the splitting, can be determined by subtracting the weighted entropy after the splitting from the entropy before splitting as in Equation (11). The feature with the highest information gain will be chosen by the algorithm as the root node of the decision tree. The formula shown in Equation (12) is used to calculate the Gini index. The value of the Gini index varies between 0 and 1, the feature with the lowest Gini-index value will be selected by the algorithm as the root node in the decision tree [30].

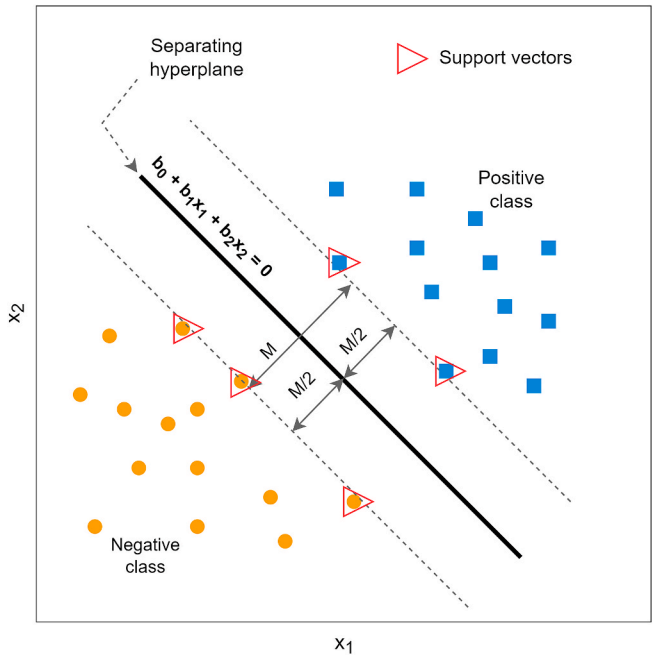

Fig. 1. SVM applied to obtain the decision boundary for the classification of two classes.

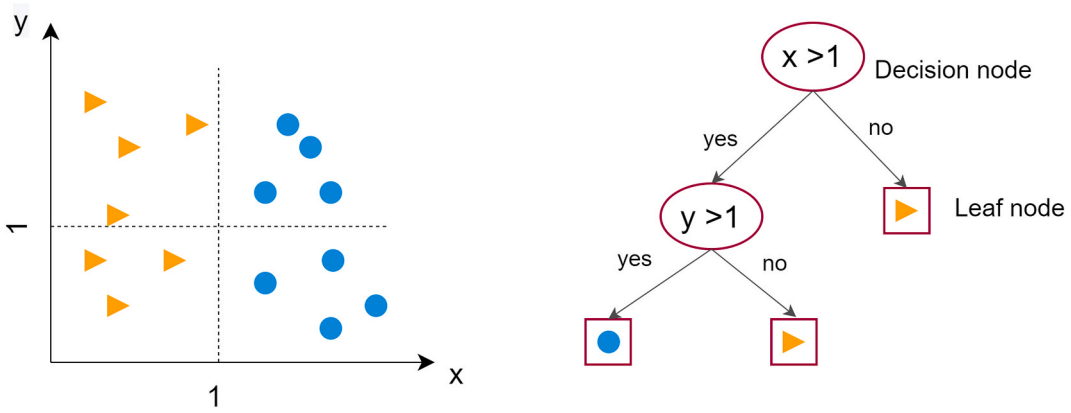

Fig. 2. An example of a dataset and its corresponding classification tree.

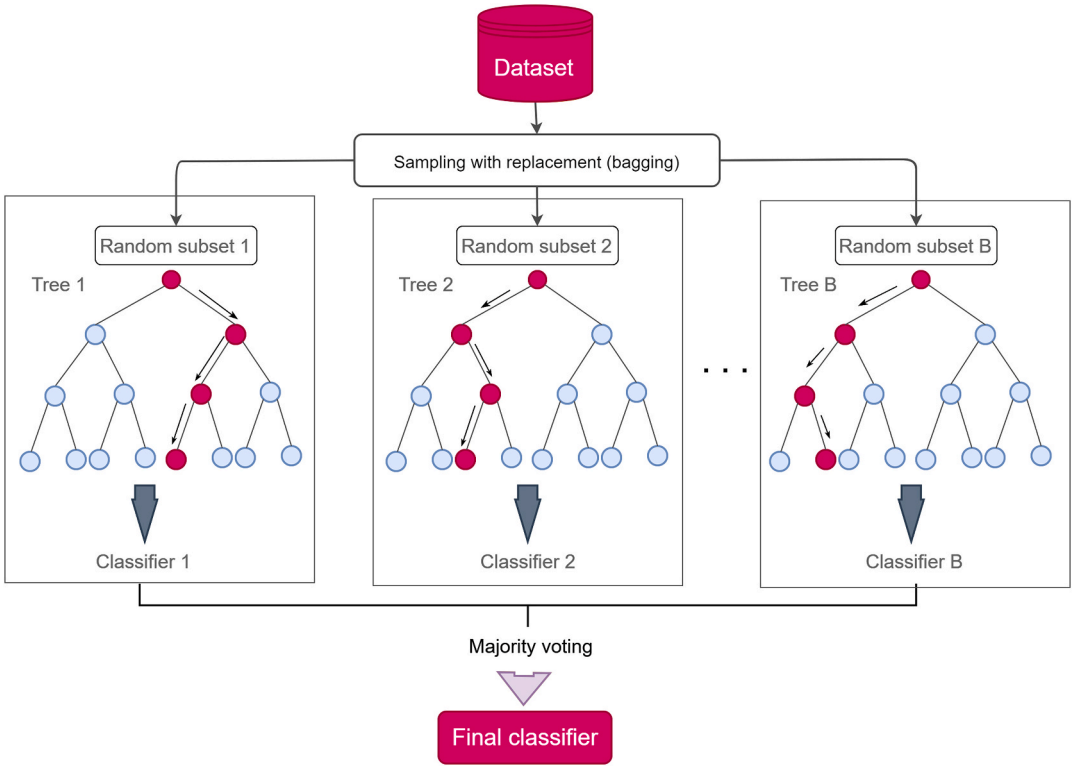

Fig. 3. A representation of the RF classification process.

$$E(S) = \sum_{i=1}^{n} -(p_i log_2 p_i) \tag{10}$$

$$IG(S,X) = E(S) - E(S|X) \tag{11}$$

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{12}$$

where p_i is the probability of the i^{th} class, $E(S)$ is the entropy of the sample dataset before splitting, $E(S|X)$ is the sum of entropies after splitting the dataset into m different classes based on X feature.

4.5. Random forests

Random forests (RF) is an ensemble algorithm that uses a collection of decision trees as the base model. It is a significant modification of bagging in which a large collection of de-correlated decision trees is built and then averaged. The basic idea behind bagging is to average many noisy but roughly unbiased models and thus reduce the variance. Trees are ideal candidates for bagging because they can capture complex interaction structures in data and have relatively low bias when grown deep enough. Additionally, trees benefit greatly from averaging since they are extremely noisy. Furthermore, since each tree generated by bagging has been identically distributed (i.d.), the expectation of an average of B such trees is the same as the expectation of any one of them. This means that the bias of bagged trees is the same as that of single trees, and the only way to improve is to reduce the variance.

An average of B i.i.d. random variables, each with variance σ^2, has variance $\frac{1}{B}\sigma^2$. If the variables are simply i.d. with positive pairwise correlation ρ, the variance of the average is shown in Equation (13).

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{13}$$

As B increases, the second term disappears but the first term persists, limiting the benefits of averaging to the magnitude of the correlation of pairs of bagged trees. The goal of random forests is to improve bagging variance reduction by reducing the correlation between trees without increasing the variance too much. This is accomplished during the tree growth process by using a random selection of input variables. When growing a tree on a bootstrapped dataset, before each split, select $m \leq p$ of the input variables at random as candidates for splitting. m is usually set to $\sqrt{p}$ or even as low as 1. After B such trees $\{T(x;\Theta_b\}_1^B$ are grown, the majority vote of each tree's output becomes the final prediction of the random forest classification model as in Equation (14) [31]. A representation of the RF classification process is illustrated in Fig. 3.

$$\widehat{C}_{rf}^{B}(x) = majority\ vote\ \{\widehat{C}_b(x)\}_1^B \tag{14}$$

where $\widehat{C}_b(x)$ is the b^{th} random-forest tree's class prediction.

5. Materials and methods

This section focuses on the experimental data used as well as the whole development process of models predicting the level of chloride resistance of concrete, from data preprocessing to model training and validation. It begins with a presentation of the experimental dataset before delving into the model development process in depth.

5.1. Experimental dataset

A database of 843 observations of non-steady-state migration coefficients (D_{nssm}) of various concrete types with the proportion of the mix components is formed by retrieving from: i) research projects of LIFECON and Finnish DuraInt-project [32], and ii) international journal articles [5], [6,7,33], [14,15,34–48]. To represent the concrete's resistance to chloride penetration, the D_{nssm} values are translated into five classes (low, moderate, high, very high, and extremely high) based on the criteria described in Table 1. Normal-weight, lightweight, high-strength, high-performance and self-compacting concrete are among the types of concrete covered in the database. The concrete mix has eight features that describe the mix components and their proportions. These are w/b ratio, contents of binders (cement, slag, FA, SF, and lime filler) in units of kg/m^3, amount of fine, coarse, and total aggregates in units of kg/m^3, contents of chemical additives (plasticizers, superplasticizers, and air-entraining agents (AEA) in % by the weight of the binder.

Table 1
Criteria for the classification of chloride penetration resistance of concrete.

D_{nssm} (x10^{-12} m^2/s)	Chloride Penetration Resistance of Concrete
>15	Low
10–15	Moderate
5–10	High
2.5–5	Very high
<2.5	Extremely high

The database contains numerous cement types described in different standards as it is based on experimental data collected in different parts of the world. For this reason, all types of cement are translated according to the European standard EN 197–1 for the sake of consistency [49]. The standard defines 27 different common types of cement, which can be divided into five groups (CEM I, II, III, IV, and V). A total of 15 types of cement from the four main cement groups are contained in the database. 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, 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). Supplementary cementitious materials are used in about 57% of the experiments. The w/b ratio ranges from 0.19 to 0.65. The admixtures in the dataset are made up of various compounds, such as superplasticizer, which is made up of lignosulfonate, melamine sulfonate, naphthalene, and polycarboxylate ether, and AEA, which is composed of fatty acid soap, synthetic surfactants, and vinsol resin. Despite the fact that admixtures behaviours differ depending on their chemical compositions, the database does not classify them based on their chemical composition. Table 2 describes the data in more detail.

5.2. Model development

The model development approach for classifying the level of resistance of concrete to chloride penetration is described in this section. Fig. 4 illustrates the pipeline of the models. Data retrieval is the first activity, followed by data preprocessing which is the most important step in building machine learning models, since the performance of the models is primarily determined by the quality of the data. A range of operations are often conducted during data preprocessing to make the data suitable for a machine learning model, including outlier detection and treatment, data encoding, feature scaling, feature engineering, and partitioning the dataset into a training and test set. The next step is to choose the appropriate algorithms and train the model with the training data. By tuning the hyperparameters, the model training process will be iterated until the best cross-validation results are obtained. The models' performance is then evaluated using a test set that the model has never seen before. The following sections cover all the important steps in the model development process.

5.3. Data preprocessing

Data preprocessing is an important stage in building any machine learning model. It includes operations such as missing data processing, outlier detection and handling, data encoding, feature scaling, feature selection, and data division. The next subsections go through all of the data preprocessing tasks used in this project.

5.3.1. Missing data processing

The quality of the data is one of the most important factors controlling the performance of any data-driven model. Data missing is a problem that impacts nearly every scientific profession. It reduces the predictive capacity of machine learning models and introduces bias in the model. As a result, the issue of missing data shall be addressed first. It can be handled in a variety of ways, including i) removing observations with any missing values, ii) relying on the learning algorithm to handle missing values during training, and iii) imputing all missing values. All missing observations are discarded in this work.

Table 2
Description of the features employed in the raw dataset.

No.	Feature subcategory		Description	Unit
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]
3	Cement content			[kg/m^3]
4	Mineral admixtures content		Slag	[kg/m^3]
5			Fly ash	[kg/m^3]
6			Silica fume	[kg/m3]
7			Lime filler	[kg/m^3]
8	Water-to-binder ratio			–
9	Aggregates content		Fine aggregate	[kg/m^3]
10			Coarse aggregate	[kg/m^3]
11			Total aggregate	[kg/m^3]
12	Chemical admixtures content		Plasticizer	[% by binder wt.]
13			Superplasticizer	[% by binder wt.]
14			Air-entraining agent	[% by binder wt.]
15	Concrete age at migration test			[days]
16	Concrete's resistance level to chloride penetration			–

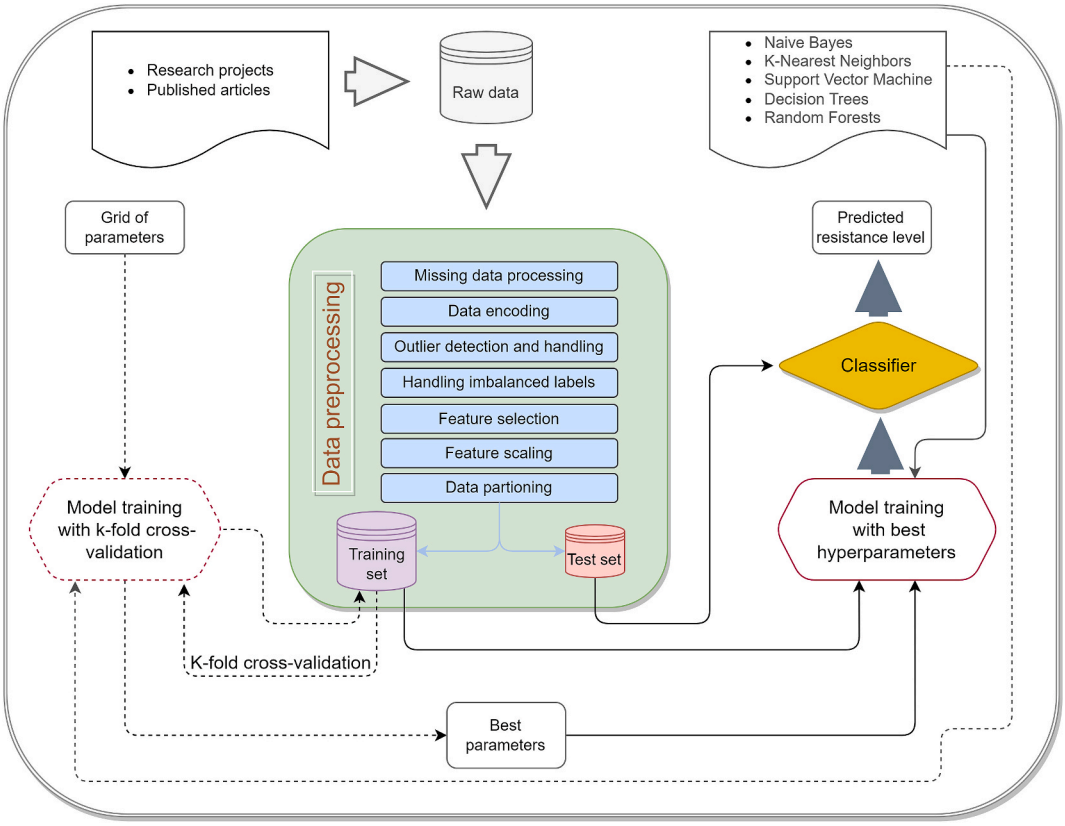

Fig. 4. The pipeline of models for classifying the degree of resistance of concrete to chloride penetration.

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

Fig. 5. One-hot encoding of the feature "cement type".

Resistance level	CEM I
Low	0
Moderate	1
High	2
Very high	3
Extremely high	4

Fig. 6. Lebel encoding of the target feature.

5.3.2. Data encoding

Many machine learning algorithms require numeric input features, so any features that involve categorical data (nonnumerical values) must be converted to numeric values. There are two features in the dataset that contain nonnumeric values. One of the features is the one that describes the cement types involved and the other is the target feature that describes the chloride resistance level of the concrete. Converting categorical values to numeric values can be done in a number of ways. Each strategy has its own trade-offs and

consequences. The feature "cement type" is converted using one-hot encoding, whereas the target feature is encoded using the label encoding method. One-hot encoding translates categorical data into a binary vector format. That is, a new column is created for each unique value in a column. This column is expressed as 1 or 0s depending on whether the value fits in the column heading. Fig. 5 presents one-hot encoding of the feature "cement type".

Label encoding techniques are used to encode the class labels of the target feature with a value between 0 and 4, as shown in Fig. 6. The numerical order does not appear to be random, which makes sense if the algorithm interprets the resistance level as $0 < 1 < 2 < 3 < 4$, that is, Low < Moderate < High < Very high < Extremely high. From this example it can be seen that there is a clear benefit to using one-hot encoding instead of label encoding for the feature "cement type". If it is converted using label encoding method, the cement type will have values ranging from 0 to 6. The machine learning algorithms may misinterpret the converted values during model training by believing that one category value is higher than the other simply because it is stored as an integer, while the data does not support such an ordinal relationship.

5.3.3. Outlier detection and handling

Outliers or observations that differ significantly from the rest of the population must be dealt with before the dataset is used to train the model. Although there are certain machine learning algorithms that are not sensitive to outliers, some of the algorithms used in this study are sensitive to outliers. In this work, one of the most effective outlier detection algorithms known as isolation forests, is utilized to isolate anomalies from the data. This method, unlike most outlier detection techniques, attempts to explicitly detect actual outliers rather than identifying normal data points that typically involve more conditions. On the other hand, isolating outliers from normal data points requires fewer conditions and is thus more efficient than isolating normal data points.

Isolation forest, like other tree-based ensembles, is construct on a collection of decision trees called isolation trees, "iTrees", with each tree containing a subset of the entire dataset. It selects n randomly selected samples of size m from a given dataset. The "iTree" is built for each random sample by splitting the subsample instances across a split value of a randomly chosen feature, so that the instances whose corresponding feature value is less than the split value go to the left and the others go to the right, and the process becomes repeated recursively until the tree is fully built. The split value is randomly chosen from the minimum and maximum values of the chosen feature. Outliers are the points with the shortest path length, $h(x)$, which is the length of the path from the root to the leaf node on each "iTrees," as illustrated in Fig. 7.

The outlier score of an instance can be computed based on the observation that the structure of "iTrees" is similar to that of Binary Search Trees (BST): the termination of the leaf node of "iTrees" corresponds to the failure of the BST search. As a result, the estimate of the mean $h(x)$ for the termination of the leaf node is the same as an unsuccessful search in BST [50,51]. That is the same as expressed in Equation (15).

$$c(m) = \begin{cases} 2H(m-1) - \dfrac{2(m-1)}{m} & \text{for } m > 2 \\ 1 & \text{for } m = 2 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $H(i)$ is the harmonic number, which can be estimated by $\ln(i) + 0.5772156649$ (Euler's constant), n is the size of the test set, m is the size of the sample set.

The value of $c(m)$ in Equation (15) is the mean $h(x)$ for a given m. It is then used to normalize $h(x)$ to obtain an estimate of the outlier score for a given instance x as given in Equation (16).

$$s(x,m) = 2^{\frac{-E(h(x))}{c(m)}} \tag{16}$$

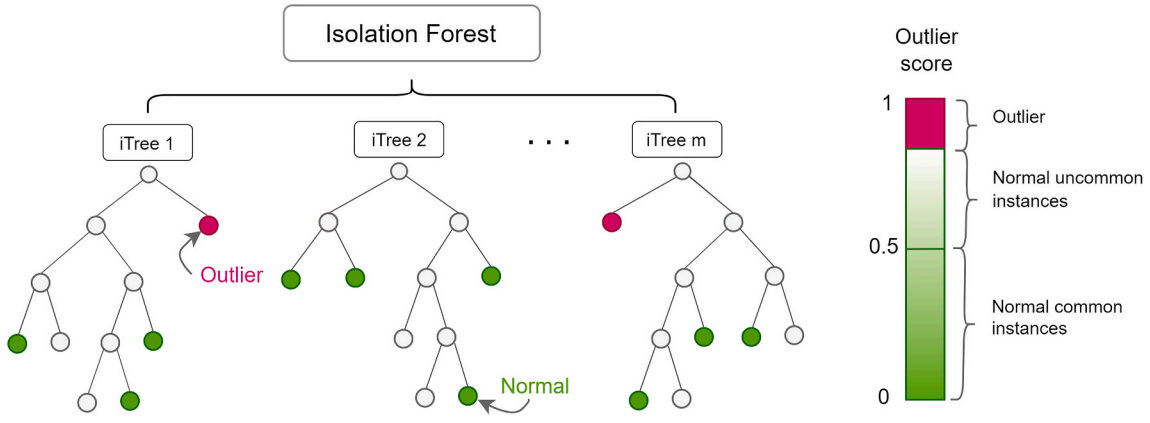

Fig. 7. Outlier detection with isolation forest.

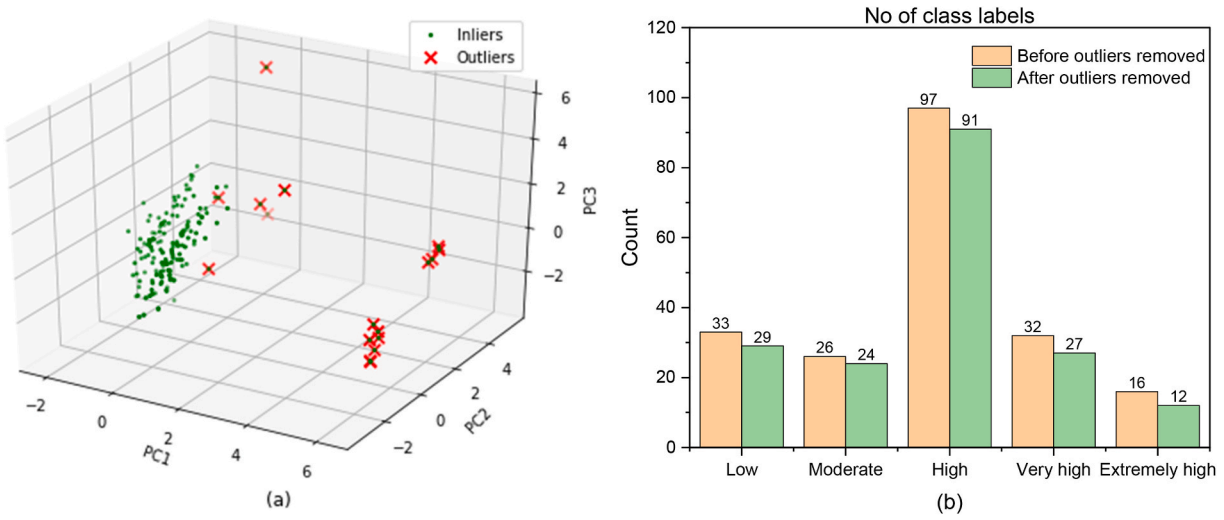

Fig. 8. Plots illustrating: (a) the detected outliers, (b) the number of outliers in each class labels.

where $E(h(x))$ is the mean $h(x)$ value from a set of isolation trees. If the value of s is close to 1, instance x is considered an outlier; if it is much smaller than 0.5, the instance x is consider normal; and if all the instances return $s \approx 0.5$, the entire sample does not have any distinct outliers.

The outliers detected by the adopted isolation forest algorithm are illustrated in Fig. 8a. It is evident to see that the outlier instances are mostly separated from the cluster of normal instances. The total number of outliers identified was 21, which is approximately 10% of the total observations and all were removed from the dataset. The outliers identified were from all class labels describing the resistance level of concrete to chloride penetration. The number of each class labels before and after removing the outliers is demonstrated in Fig. 8b.

5.3.4. Handling class imbalanced

Imbalanced datasets or datasets with imbalanced classes are those in which one target class occurs far more frequently than the other. In many domains, having imbalanced data is the norm and one of the most difficult problems in machine learning classification tasks. Because the classifier model should not be geared towards only recognizing the majority class but should also give equal importance to the minority class. The dataset used in this study is an imbalanced dataset due to the uneven distribution of the target classes that describe the concrete's level of resistance to chloride ion penetration. For instance, there is a 40% difference between the most frequent class 'High' and the least frequent class 'Extremely high'. There are different types of techniques to handle imbalanced data. In this work, the Synthetic Minority Oversampling Technique or SMOTE is applied to address the problem of the imbalanced classes. This technique oversamples the minority class by simply generating data points on the line segment connecting a randomly

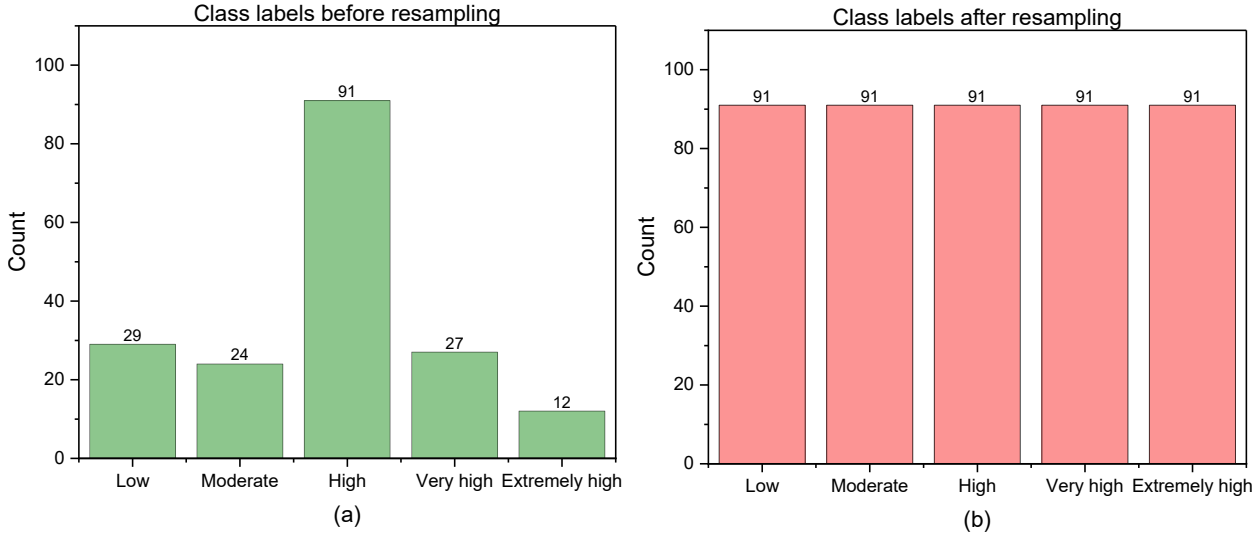

Fig. 9. Distribution of the class labels: (a) before resampling, (b) after resampling.

chosen data point and one of its K-nearest neighbors [52]. Since this approach is very simple and extremely effective in practice, it has become widespread. Fig. 9 illustrates the distribution of the class labels before and after resampling.

5.3.5. Feature selection

Feature selection is a crucial step in the data preprocessing phase for identifying important features and removing irrelevant or redundant ones. It decreases dimensionality while improving prediction performance and model training efficiency. Irrelevant features must be discarded as they affect model accuracy and cause model training to slow down. Some features may not be predictive, or they may be redundant with other feature. There are three types of feature selection approaches that can be used to help select the best features for model training. These are filter, wrapper, and embedded [53]. In this work, the embedded technique based on an RF algorithm is used to select relevant features from the dataset, since it combines the benefits of the filter and the wrapper approaches in terms of minimal computational effort and adequate accuracy. The RF used measures the importance of a feature as the averaged impurity decrease derived from all decision trees in the forest, without assuming anything about the data.

Feature importance measured by an RF algorithm is illustrated in Fig. 10. The measures of the feature importance added up to one. Binders (cement, fly ash, slag, and silica fume) collectively have the highest predictive power of the model, accounting for about 32%, followed by aggregates (fine, coarse, and total aggregate) at approximately 29%. The features w/b ratio, migration test age, and water account for about 13, 11 and 8%, respectively. The feature cement types are considered to be powerless predictors for the classification of concrete chloride resistance for this specific dataset.

5.3.6. Feature scaling

Normalization and standardization are two approaches to scaling features onto the same scale. Even if normalization is a commonly used technique for obtaining values in a bounded interval, standardization can be more practical for many machine learning algorithms, particularly optimization algorithms like gradient descent. This is because many linear models set the weights to zero or small random values close to zero. Such algorithms may perform poorly when the individual features do not resemble to standard normally distributed data. Since linear algorithms such as KNN and SVM are used in this work, a standardization technique is used to scale the features. In this method, the mean of each feature is centered at zero, and each feature has a standard deviation of one, resulting in the property of a standard normal distribution, which makes learning the weights easier. Fig. 11 shows the feature "Cement content" before and after feature scaling as an example. The standardization procedure can be expressed mathematically as in Equation (17).

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x} \tag{17}$$

where, μ_x represents the sample mean of a specific feature and σ_x represents the corresponding standard deviation.

5.3.7. Data partitioning

Training/test partitioning usually involves splitting the data into a training and a test set in a predetermined ratio. The training set is used to train the model, whereas the test set is used to evaluate the performance of the trained models against data that has never seen before. In this work, the data are randomly divided into two parts: 80 and 20%. 80% of the data is allocated for model training and the remaining 20% for testing the model's performance.

5.4. Model training and evaluation

In this work, two scenarios (Scenario 1 and 2) for the development of the models are considered based on the input features. Scenario 1 uses all the features presented in Table 3, whereas Scenario 2 takes into account all features considered in Scenario 1, except for the features that describe the cement types. This feature was considered as powerless by the applied feature section method. The aim of Scenario 2 is to investigate the influence of feature selection on the classification accuracy of the chloride resistance of concrete.

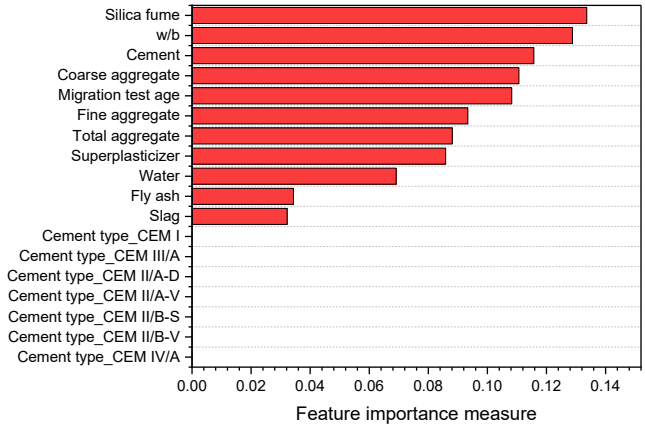

Fig. 10. The feature importance measure of the mix ingredients and concrete age at chloride migration test.

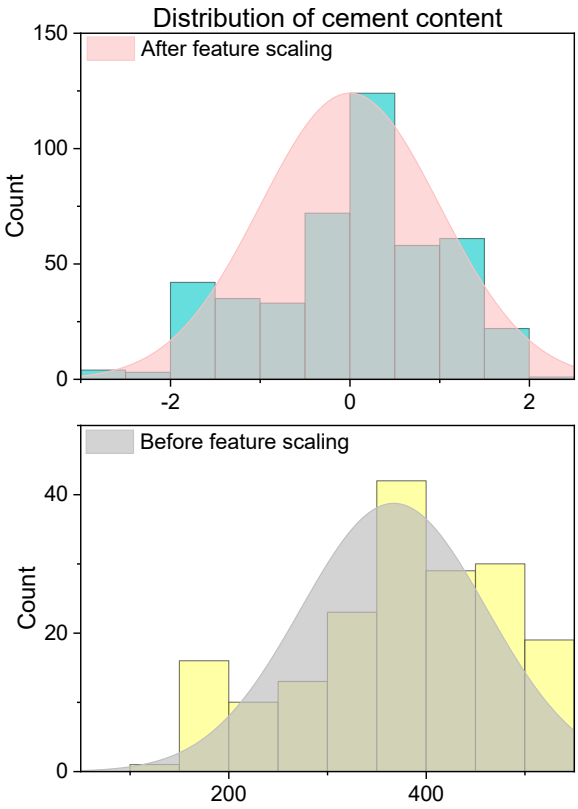

Fig. 11. Distribution of the feature "Cement content" before and after feature scaling.

Table 3
Input features considered for the development of chloride resistance classifier models.

No.	Feature subcategory		Description	Scenario 1	Scenario 2
1	Cement types	CEM I	CEM I	✓	×
		CEM II	CEM II/B–S, CEM II/A-D, CEM II/A-V, CEM II/B–V	✓	×
		CEM III	CEM III/A	✓	×
		CEM IV	CEM IV/A	✓	×
2	Water content			✓	✓
3	Cement content			✓	✓
4	Mineral admixtures content		Slag	✓	✓
5			Fly ash	✓	✓
6			Silica fume	✓	✓
7	Water-to-binder ratio			✓	✓
8	Aggregates content		Fine aggregate	✓	✓
9			Coarse aggregate	✓	✓
10			Total aggregate	✓	✓
11	Chemical admixtures content		Superplasticizer	✓	✓
12	Concrete age at migration test			✓	✓

× = not applicable, ✓ applicable.

Compared to the raw data presented in Table 2, three features are missing in Table 3, namely air-entraining, plasticizer, and lime filler. This is because these features obtained null values after all missing values are eliminated from the dataset during the data pre-processing phase.

Descriptive statistics of the numerical features of the preprocessed data that are used to train and validate the models is presented in Table 4. It is worth noting that the final dataset contains 204 observations. The w/b of the concrete in this database ranged from 0.3 to 0.65. The cement content ranges between 52 and 525 kg/m^3. The amount of supplementary cementitious materials of slag, fly ash, and silica fume used to partially replace cement reached up to 312.30 kg/m^3, 735 kg/m^3, and 468.50 kg/m^3, respectively. The wide range of ingredient amounts confirms the inclusion of various types of concrete in the dataset.

A total of ten models, five in each scenario, using the NB, KNN, SVM, DT, and RF are trained using the training dataset. In order to obtain high-performing models, the hyperparameters of all the algorithms are tuned using a grid search method that performs an

Table 4
Descriptive statistics of the dataset.

	w/b [−]	Water [kg/m^3]	Cement [kg/m^3]	Slag [kg/m^3]	Fly ash [kg/m^3]	Silica fume [kg/m^3]	Fine aggregate [kg/m^3]	Coarse aggregate [kg/m^3]	Total aggregate [kg/m^3]	Superplasticizer [% by binder wt.]	Migration test age [days]
count	204	204	204	204	204	204	204	204	204	204	204
mean	0.42	175.99	361.85	18.44	44.12	8.36	798.83	797.21	1596.04	0.41	65.57
std	0.08	22.64	96.90	57.35	133.70	35.46	223.80	306.18	291.55	0.54	83.70
min	0.30	122.50	52.00	0.00	0.00	0.00	235.00	0.00	630.00	0.00	3.00
25%	0.36	158.00	297.88	0.00	0.00	0.00	685.26	451.50	1415.82	0.00	28.00
50%	0.40	175.00	350.00	0.00	0.00	0.00	765.00	915.53	1720.00	0.20	28.00
75%	0.45	191.00	444.38	0.00	0.00	0.00	956.75	1059.95	1801.00	0.70	90.00
max	0.65	222.00	525.00	312.30	735.00	468.50	1574.10	1240.00	1950.00	4.17	365.00

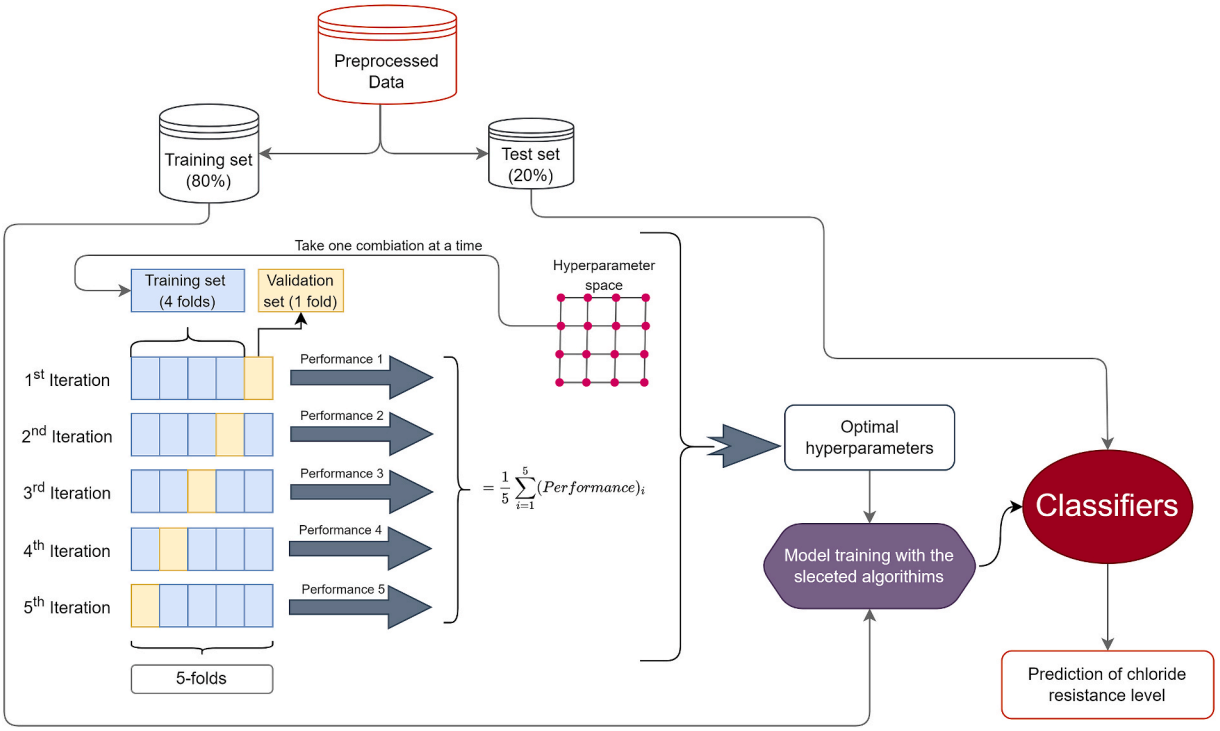

Fig. 12. A grid search accompanied by a 5-fold cross-validation method.

Table 5
Hyperparameters considered in the model development and the identified optimal hyperparameters.

Learning algorithms	Hyperparameters	Grid search ranges	Optimal hyperparameters		Description of the hyperparameters
			Scenario-1	Scenario-2	
Naïve Bayes	var_smoothing	[1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, 1e-10, 1e-11, 1e-12, 1e-13, 1e-14, 1e-15]	0.01	0.01	Portion of the largest variance of all features that is added to variances for calculation stability.
k-Nearest Neighbors	n_neighbors	1 to 40	1	1	Number of neighbors.
	weights	['uniform','distance']	'uniform'	'uniform'	Weight function used in prediction
Decision Trees	criterion	['entropy', 'gini']	'entropy'	'entropy'	The function to measure the quality of a split.
	ccp_alpha	[0.1, 0.01, 0.001]	0.001	0.001	Complexity parameter used for Minimal Cost-Complexity Pruning.
	max_depth	['none', 5, 6, 7, 8, 9, 10, 20, 40, 100, 200, 500]	500	500	The maximum depth of the tree.
	min_samples_split	[2,5,10]	2	5	The minimum number of samples required to split an internal node
	min_samples_leaf	[1–6]	1	3	The minimum number of samples required to be at a leaf node
Support Vector Machine	kernel	['rbf', 'linear']	'rbf'	'rbf'	Specifies the kernel type to be used in the algorithm.
	'C'	[0.001, 0.01, 0.1, 1, 10, 100]	10	100	Regularization parameter.
	'gamma'	[0.001, 0.01, 0.1, 1, 10, 100]	0.1	0.1	Kernel coefficient.
Random Forests	'max_depth'	[2,4–6]	6	6	The maximum depth of the tree.
	'max_features':	[1–4]	4	4	The number of features to consider when looking for the best split.
	'min_samples_leaf'	[3–5]	4	4	The minimum number of samples required to be at a leaf node.
	'min_samples_split'	[4,6,8,10]	4	6	The minimum number of samples required to split an internal node.
	'criterion'	['gini', 'entropy']	'entropy'	'entropy'	The function to measure the quality of a split.
	'n_estimators'	[10, 50, 100]	50	100	The number of trees in the forest.

exhaustive search through a manually defined subset of the hyperparameter space of a learning system. It is often accompanied by a cross-validation or hold-out performance metrics. The K-fold cross-validation approach is used in this study. Using this method, the training dataset was randomly partitioned into K subgroups of nearly 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. K models are fitted in total, obtaining K validation evaluation metrics. The overall performance of the model is determined by averaging the scores of the K-folds.

In this work, a grid search method with 5-fold cross-validation is used. This procedure is illustrated in Fig. 12. A total of ten models (five from each scenario) were trained using the five algorithms. The hyperparameters considered for each learning algorithms along with their grid search ranges are presented in Table 5. The same table also lists the best hyperparameters that can provide high prediction accuracy for all algorithms. In both scenarios, about 70% of the determined optimal hyperparameters are identical. Following the determination of the best hyperparameters for each algorithm, the identified optimal hyperparameters are used to train the data and build models that classify the degree of resistance of concrete to chloride penetration.

After the models have been trained with the optimal hyperparameters, their predictive abilities are to be evaluated with unseen data (test dataset). The classification performance metrics, confusion matrix, accuracy, precision, recall, F1-score, and area under the receiver operating characteristic are used to assess the performance of the models. The descriptions of these performance evaluation metrics are presented in the following subsections.

5.4.1. Confusion matrix

A confusion matrix, also known as an error matrix, is an x-by-x array that summarizes a classification model's performance, where x is the number of classes. The number of correct and incorrect predictions is summarized with count values and divided by each class. Fig. 13 shows an example of a confusion matrix of a binary classification (negative and positive class). It represents combinations of actual and predicted values. The rows represent the true classes, while the columns represent the predicted classes. Entries on the main diagonal of the confusion matrix correspond to correct classifications, while other entries indicate how many samples of one class were incorrectly classified as samples of another class. The key terms in the confusion matrix entries are true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP values are those that were actually positive and were predicted to be positive. FP values are those that were predicted to be positive but were actually negative. FN values are those that were actually positive but were incorrectly predicted to be negative. TN values are those that were predicted to be negative and were actually negative.

Although the confusion matrix provides useful information about the performance of the classification model, examining the entire confusion matrix takes time because the assessment process is very manual and qualitative. The information in the confusion matrix can be aggregated in a variety of ways and used as metrics to evaluate classifier performance. These are accuracy, precision, recall, and F1-score.

5.4.2. Accuracy

Accuracy is a performance metric that express the overall performance of a classifier as the number of correct predictions (TP and TN) divided by the total number of all samples. It demonstrates the ability of the classification models to correctly classify data samples. It is calculated as shown in Equation (18).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

5.4.3. Precision and recall

Precision, also known as positive predictive value, is a measure of how many of the samples predicted to be positive are actually

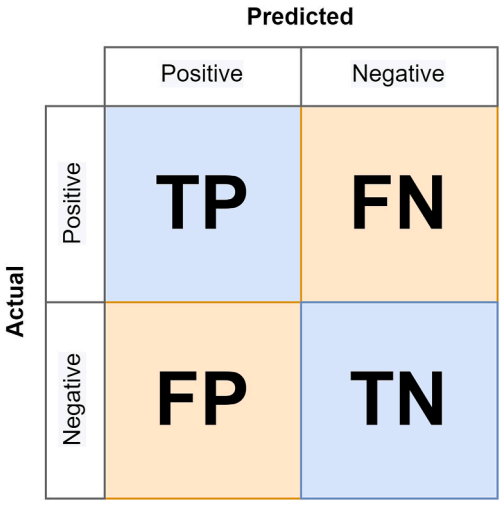

Fig. 13. An example of a binary classification confusion matrix.

positive. It is used as a performance metric when the goal is to limit the number of false positives. Precision is calculated as shown in Equation (19).

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

Recall, also known as sensitivity, hit rate probability of detection, or true positive rate (TPR), on the other hand, is a measure of how many positive samples are captured by the positive predictions. It is used as a performance metric when identification of all positive samples is required. In other words, when it comes to avoiding false negatives. The formula shown in Equation (20) is used to calculate recall.

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

Precision and recall are two of the most widely used binary classification measures. While both are important metrics, focusing on just one will not provide a complete picture. The F1-score is one way to summarize them.

5.4.4. F1-score

F1-score is the harmonic mean of precision and recall. It can be a better measure than accuracy in imbalanced binary classification datasets because it accounts for precision and recall. The F1-score is calculated as in Equation (21).

$$\mathrm{F1-score} = 2^{*}\frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{21}$$

In the case of multiclass, in addition to calculating the above-mentioned metrics per class label, it is crucial to compute their average using one of the following strategies. i) macro average: calculates the unweighted average across classes. This will give all classes the same weight regardless of their size, and ii) weighted average: calculates the average between classes weighted by the number of observations in each class. For instance, the macro average and weighted average of the precision score in a k-class system can be calculated from the system's individual TPs, TNs, FPs, and FNs of shown in Equations (22) and (23).

$$Macro\ average\ precision = \frac{\sum_{i=1}^{k} Precision_i}{k} \tag{22}$$

$$Weighted-average\ precision = \frac{\sum_{i=1}^{k} Precision_i * N_i}{\sum_{i=1}^{k} N_i} \tag{23}$$

where k is the number of classes, N_i is the number of observations in class i, and $Precision_i$ is the precision score of class. i .

5.4.5. Area under the receiver operating characteristic

The area under the receiver operating characteristic (AUROC) is a common performance metric for evaluating binary classifiers. The receiver operating characteristic (ROC) is a graph plotting the TPR versus the false positive rate (FPR), which are calculated by shifting the classifier's decision threshold. FRP is computed as $FP/FP + TN$. The area under the curve (AUC) is the area beneath the ROC curve. It can be viewed as the probability that the model will rank a random positive example higher than a random negative sample. AUROC can also be applied to multiclass classifiers using one of two commonly used strategies: one-vs-one (OvO) and one-vs-rest (OvR). The former calculates the average of the pairwise AUROC values and the latter computes the average of the AUROC scores for each class against all other classes.

6. Results and discussion

The performance of all ten trained models under two scenarios (five models each) is presented and discussed in this section. NB, KNN, SVM, DT, and RF were the learning algorithms used in both scenarios. The main distinction between the two scenarios is the number of input features describing the concrete mix's ingredients. The first scenario takes into account all of the features described in

Table 6
Performance metrics of the developed models.

	Algorithm	Precision	Recall	F1-score	Accuracy	AUROC
Scenario 1	NB	0.67	0.66	0.65	0.66	0.9077
	KNN	0.88	0.88	0.88	0.88	0.9232
	DT	0.86	0.86	0.86	0.86	0.9096
	SVM	0.89	0.88	0.88	0.88	0.9829
	RF	0.85	0.85	0.85	0.85	0.9792
Scenario 2	NB	0.62	0.60	0.59	0.60	0.8687
	KNN	0.87	0.87	0.86	0.87	0.9166
	DT	0.83	0.81	0.82	0.81	0.9161
	SVM	0.89	0.89	0.89	0.89	0.9714
	RF	0.83	0.82	0.82	0.82	0.9627

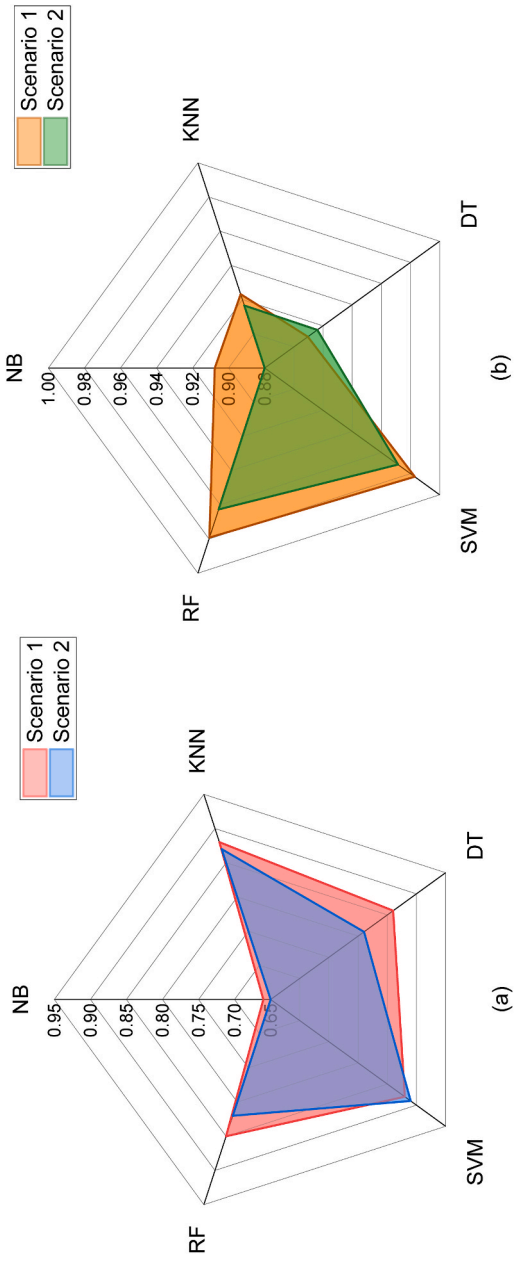

Fig. 14. Performance of the developed models: (a) accuracy, (b) AUROC.

Table 3, whereas the second scenario considers only a subset of the features.

All trained models are validated using the test dataset derived from the original data but not included in the training dataset. A detailed analysis of the performance of all models extracted from the confusion matrices (precision, recall, F1-score, and accuracy) as well as one-vs-the-rest AUROC is presented in Table 6. All of these matrices are the weighted averages. As in this work the imbalanced data was resolved prior to model training, evaluating the performance of the models with their accuracy score is reliable. According to the results shown in the table, all of the learning algorithms except NB predict chloride resistance to chloride penetration with an accuracy better than 80%. Though the performance of most models is fairly comparable, SVM models from Scenario 2 outperform all others with an accuracy score of 0.89 and an AUROC of 0.9714. This means that the model has an accuracy of 89% and predicts the classes with a probability of approximately 97.14%. All of the assessment metrics (precision, recall, and F1-score) of this algorithm are also higher than any of the models. The Scenario 1 KNN and SVM models have an accuracy score of 0.88. Between the two models, SVM had the highest AUROC at 0.9829 followed by KNN at 0.96232. This confirms that SVM is the superior model in Scenario 1 as well. NB models perform worst in both scenarios, with an accuracy of 66% in the case of Scenario 1 and 60% in the case of Scenario 2. The accuracy retained by most of the models is remarkably high considering the fact that the dataset comprised limited observations covering a wider range of concrete types.

The accuracy score and AUROC of all models are shown in Fig. 14 to give a general overview of their performance. It is clear that almost all of the models in Scenario 1 are slightly superior to those in Scenario 2 as can be seen from the accuracy plot in Fig. 14a. Only SVM from Scenario 2 slightly outperform the corresponding models from Scenario 1. In comparison to the corresponding models in Scenario 1, this model from Scenario 2 classifies the chloride resistance level of concrete with a less degree of certainty as show in Fig. 14b. Indeed, the performance of the models may vary slightly between runs as the models trained with randomly selected data produces a different model with different capabilities. Despite this fact, the models in Scenario 1 generally perform slightly better than the models in Scenario 2. This corroborates that the inclusion of all features that describe the mix components is essential for a more accurate and confident classification of the chloride resistance of concrete. Although considering only the influential features identified by the adopted RF-based feature selection method, as in Scenario 2, does not improve prediction performance, all of the models in this scenario are efficient and parsimonious almost without sacrificing prediction accuracy. It is worth noting that this study uses only five types of commonly used algorithms to classify the chloride resistance of concrete. It is necessary to examine other learning algorithms to determine if there are other algorithms that may be better suited to address this problem.

The true and predicted labels describing the level of chloride resistance to chloride penetration are shown in Fig. 15 as confusion matrices. The figure on the left shows the confusion matrix of the best performing model in Scenario 1, while the figure on the right is in Scenario 2. The difference in the number of correctly classified and misclassified labels between the models found in the two models shows almost the same trend. For example, the SVM algorithm in Scenario 1 correctly classifies the chloride resistance of concrete as low, moderate, high, very high, and extremely high in 16/16 (100%), 16/22 (73%), 16/20 (80%), 12/13 (92%), and 20/20 (100%) cases. In the case of Scenario 2, the same algorithm correctly classifies the chloride resistance levels (same order as above) as 17/17 (100%), 17/20 (85%), 16/21 (76%), 14/16 (88%), and 17/17 (100%). Though the SVM of Scenario 2 outperforms that of in Scenario 1, this does not imply that the algorithm is better at predicting each class label. For instance, it predicts better for the labels “High” and “Very high” in the case of Scenario 1. The same phenomenon is also observed in the other algorithms.

The ROC curves of the SVM models from the two Scenarios are plotted in Fig. 16. A perfect classifier would be in the top left corner of the chart, with a TPR of 1 and an FPR of 0. In other words, the greater the area under the curve (the closer the AUROC is to 1), the better model. The diagonal of an ROC plot can be interpreted as random guessing, and classification models that fall below the diagonal (AUROC less than 0.5) are considered worse than random guessing. Fig. 16 shows that all ROCs are well above the diagonal line, showing the quality of the models. In fact, AUC for each class label is also shown in the chart. It is worth noting that the AUC for all class labels except class 2 (resistance level “High”) is either 0.98, 0.99 or 1. This confirms that the models predict the chloride resistance level of concrete with a high degree of certainty.

The validity of all models was tested on unseen data, and the results demonstrated their highly accurate predictive ability. The models classify the level of concrete resistance to chloride penetration with high degree of certainty, demonstrating their high generalizability. All these facts confirm that models accurately capture the complex interactions among the interacting features, and thus the research findings of this work are reliable and valid. The high accuracy of the trained machine learning based models also confirms their practicability for predicting the level of concrete resistance to chloride penetration. As the chloride diffusivity is the primary durability indicator, the models could help provide a quick overview of the durability and service life of reinforced concrete building structures exposed to coastal or chloride-loaded environments. Furthermore, the models greatly aid building/civil engineers in designing concrete that can withstand the desired level of chloride penetration. All of these facts have significant economic implications for society as the models help define proactive maintenance, which ultimately reduces lifecycle costs considerably. The models also aid in the design of sustainable concrete mixes capable of withstanding the desired level of chloride penetration or in obtaining quick results on newly designed concrete without the need for time-consuming and resource-intensive laboratory testing. It is worth noting that the models were developed using a variety type of concrete mixes from around the world, which represent common concrete mixes of many countries. As a result, they could be universally applicable to predict the chloride resistance of concrete and thus aid in the assessment of the durability and service life of coastal concrete buildings or buildings exposed to chloride-containing deicing salts.

The performance of the model could be further improved with more representative data. Obtaining data from published studies, on the other hand, was a difficult task. The use of multiple measurement units, the absence of certain features, and the various methods of providing information were all common issues. This is because individual studies performed in different countries only consider a limited number of features. To produce well-rounded data, all of the data had to be translated into appropriate units and formats,

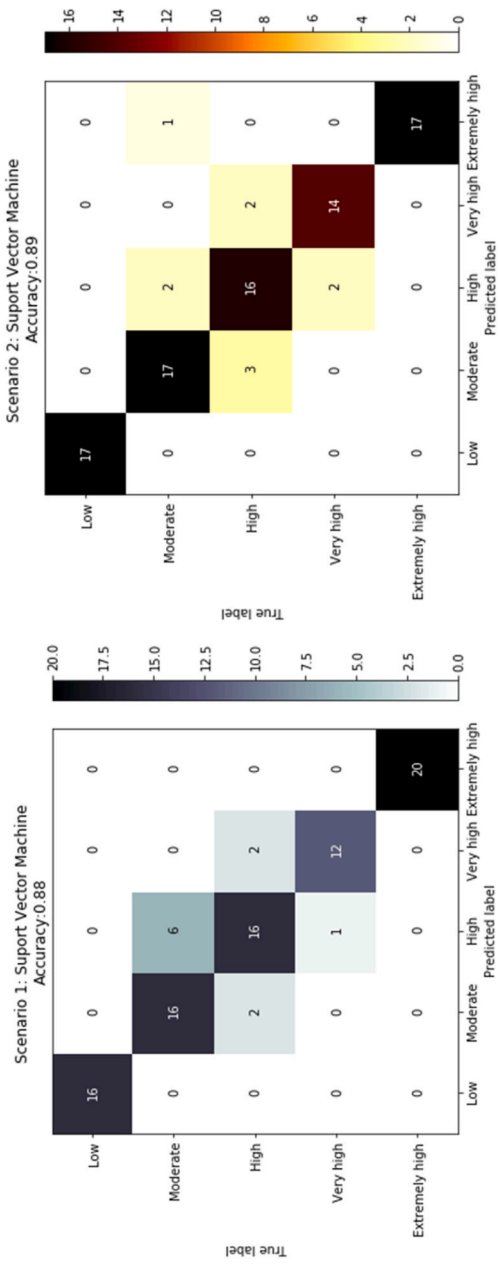

Fig. 15. Confusion matrices of the best performing models in Scenarios 1 and 2.

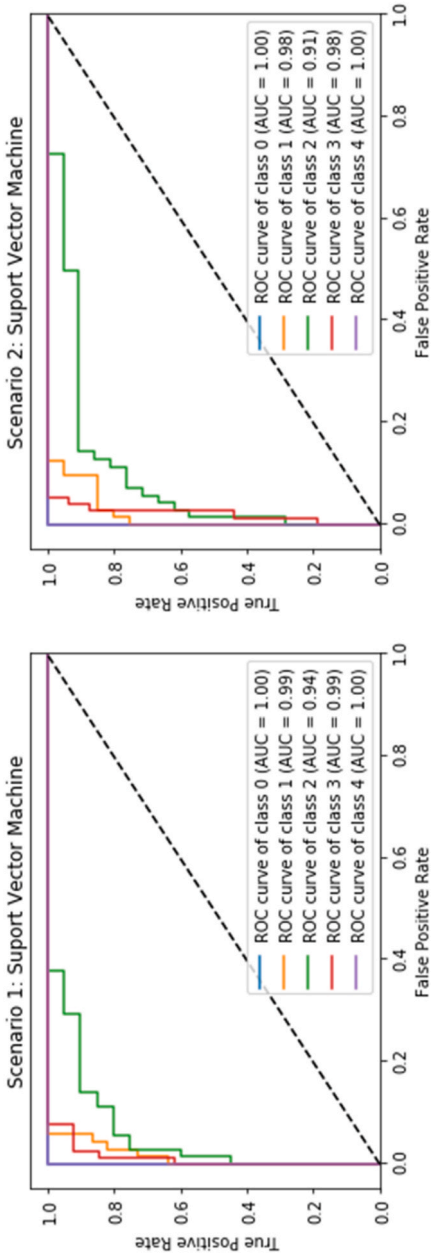

Fig. 16. ROC curve of the two best performing models with AUROC.

which took a long time and a lot of 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. Indeed, as sensing technology advances, the availability of miniaturized, stable, and low-cost sensors for monitoring the durability of concrete structures could increase in the coming years, resulting in an influx of data [54,55]. Hence in order to reap significant benefits from machine learning, open data in a machine-readable format that can be freely shared among domain experts is required. Ultimately, the building and infrastructure engineering sector will no longer be an outlier in the ubiquitous digital revolution.

All of the models presented in this work were formulated by writing codes in a Jupyter notebook environment using Python, a popular open-source programming language. Jupyter Notebook is a simple to use open-source web application that provides an interactive computational environment for creating, executing, and visualizing interactive data in a variety of programming languages. Building/civil engineers familiar with Python programming and Jupyter notebook can use the developed models to assess concrete resistance to chloride penetration and thus get a quick overview of the durability and service life of coastal reinforced concrete buildings or buildings exposed to chloride-containing deicing salts. Indeed, because Jupyter notebook is easy to use, engineers familiar with the Python programming language can quickly learn the notebook and apply the models. In fact, the development of a user interface is critical to making the models easier to use for anyone with no prior knowledge of the Python programming language, and this is something that is planned for the future.

7. Conclusions

This study adopted five machine learning algorithms, naïve bayes, k-nearest neighbors, decision trees, support vector machine, and random forests to predict the chloride resistance level of concrete considering two scenarios. The first scenario considered all features that describe the concrete mixture components, whereas the second scenario considered only a subset of the features identified as important by the RF-based feature selection method. A dataset of concrete mix components of different concrete types with their chloride resistance values was used for the training and validation of the models. The following are the main conclusions drawn from this work:

- Performance: The validation results confirmed that all models except NB classified the level of chloride resistance of concrete with remarkably high accuracy. Among all algorithms, the SVM performed the best, with 89% and 88% accuracy in Scenarios 1 and 2, respectively. Despite this, the models in Scenario 1 performed slightly better than the models in Scenario 2. Although just taking the influential features into account, as in Scenario 2, does not improve prediction performance, all of the models in this scenario were efficient and parsimonious almost without sacrificing prediction accuracy.
- Implication: Because chloride diffusivity is the primary durability indicator, the models' high accuracy confirms their applicability for predicting the level of concrete resistance to chloride penetration and thus help provide a quick overview of the durability and service life of reinforced concrete building structures exposed to coastal or chloride-loaded environments. In addition, the models greatly aid building/civil engineers in designing concrete that can withstand the desired level of chloride penetration without the need for time-consuming and resource-intensive laboratory testing. All of these facts have significant economic implications for society.
- Universality: As all models were created using a variety of concrete mixes from around the world, which represent common concrete mixes in many countries, they can be used universally to predict the chloride resistance of concrete, helping to assess the durability and service life of reinforced concrete building structures subjected to coastal or chloride-loaded environments.
- Accessibility and reusability: All of the models are easily and freely accessible to anyone because they were created using Python, an open-source programming language, in a Jupyter notebook environment, which is also an open-source web application. Building/civil engineers who are familiar with Python programming can easily use the models to assess the degree of resistance to chloride penetration of concrete and hence its durability. All models can be reproduced with little or no effort.
- Open data: An open data exchange platform that allows the scientific community and/or the concrete community to exchange data in a standard format is needed as it takes a long time to produce well-rounded data covering a wide range of concrete types from previously published work. This will ultimately favor the building and infrastructure engineering sectors to reap the significant benefits of the ubiquitous digital revolution.

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.jobe.2022.105146.

References

[1] United Nations, Factsheet: People and Oceans, 2017.
[2] J. Liu, G. Ou, Q. Qiu, F. Xing, K. Tang, J. Zeng, Atmospheric chloride deposition in field concrete at coastal region, Constr. Build. Mater. 190 (2018) 1015–1022, https://doi.org/10.1016/j.conbuildmat.2018.09.094.
[3] C. Houska, Deicing Salt – Recognizing the Corrosion Threat, 2009, pp. 1–11.
[4] W.Z. Taffese, Data-driven Method for Enhanced Corrosion Assessment of Reinforced Concrete Structures, University of Turku, 2020. https://www.utupub.fi/handle/10024/149752.
[5] 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.
[6] 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.
[7] 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.
[8] Life-365TM Consortium III, Life-365^{TM} Service Life Prediction Model^{TM} Version and Computer Program for Predicting the Service Life and Life-Cycle Cost of Reinforced Concrete Exposed to Chlorides, 2020.
[9] DuraCrete, DuraCrete Final Technical Report: Probabilistic Performance Based Durability Design of Concrete Structures, 2000.
[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] NT BUILD 443, Concrete, Hardened: Accelerated Chloride Penetration, NORDTEST, 2010.
[12] NT BUILD 492, Concrete, Mortar and Cement-Based Repair Materials: Chloride Migration Coefficient from Non-steady-state Migration Experiments, NORDTEST, 1999.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] S. Marsland, Machine Learning: an Algorithmic Perspective, Chapman and Hall/CRC, Boca Raton, FL, 2011.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] S. Ben-David Shai Shalev-Shwartz, Understanding Machine Learning: from Theory to Algorithms, Cambridge University Press, New York, NY, 2014.
[29] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, FL, 2009, https://doi.org/10.1201/9781420059496. First.
[30] P. Cichosz, Data Mining Algorithms: Explained Using R, John Wiley & Sons, Ltd, 2015, https://doi.org/10.1002/9781118950951.
[31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second, Springer, New York, NY, 2009, https://doi.org/10.1007/978-0-387-84858-7.
[32] H. Kuosa, Concrete Durability Field Testing in DuraInt-Project: Field and Laboratory Results 2007 - 2010, Espoo, 2011.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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.

[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] 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.
[48] 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. (2016) 11, https://doi.org/10.1155/2016/2042918, 2016.
[49] EN 197-1, Cement- Part 1: Composition, Specifications and Conformity Criteria for Common Cements, CEN, 2011.
[50] F.T. Liu, K.M. Ting, Z.H. Zhou, Isolation forest, in: 2008 Eighth IEEE Int. Conf. Data Min., IEEE, 2008, pp. 413–422, https://doi.org/10.1109/ICDM.2008.17.
[51] R.C. Ripan, I.H. Sarker, M.M. Anwar, M.H. Furhad, F. Rahat, M.M. Hoque, M. Sarfraz, An isolation forest learning based outlier detection approach for effectively classifying cyber anomalies, in: A. Abraham, T. Hanne, O. Castillo, N. Gandhi, T. Nogueira Rios, T. Hong (Eds.), Hybrid Intell. Syst., Springer, Cham, 2021, pp. 270–279, https://doi.org/10.1007/978-3-030-73050-5_27.
[52] D. Elreedy, A.F. Atiya, A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance, Inf. Sci. (Ny) 505 (2019) 32–64, https://doi.org/10.1016/j.ins.2019.07.070.
[53] 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.
[54] 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.
[55] 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.