Research Article

HaoYang Huang*, Muhammad Nasir Amin*, Suleman Ayub Khan*, Kaffayatullah Khan, and Muhammad Tahir Qadir

Efficacy of sustainable cementitious materials on concrete porosity for enhancing the durability of building materials

https://doi.org/10.1515/rams-2024-0056
received April 22, 2024; accepted September 18, 2024

Abstract: The degradation of concrete structures is significantly influenced by water penetration since water serves as the primary vehicle for the movement of harmful compounds. The process of capillary water absorption is widely recognized as a crucial indicator of durability for unsaturated concrete, as it allows dangerous substances to enter the composite material. The water absorption capacity of concrete is intricately linked to its pore structure, as concrete is inherently porous. The main goal of this work is to create an innovative predictive tool that assesses the porosity of concrete by analyzing its components using a machine-learning (ML) framework. Seven distinct batch design variables were included in the generated database: fly ash, superplasticizer, water-to-binder ratio, curing time, ground granulated blast furnace slag, binder, and coarse-to-fine aggregate ratio. Four distant ML algorithms, including AdaBoost, linear regression (LR), decision tree (DT), and support vector machine (SVM), are utilized to infer the generalization capabilities of ML algorithms to estimate concrete porosity accurately. The RReliefF algorithm was implemented to calculate the significant features influencing porosity. This study concludes that in comparison to the alternative techniques, the AdaBoost method demonstrated superior performance with an R^2 score of 0.914, followed by SVM (0.870), DT (0.838), and LR (0.763). The results of the evaluation of RReliefF indicated that the binder possesses a remarkable influence on the porosity of concrete.

Keywords: concrete porosity, durability, supplementary cementitious materials

*** Corresponding author: HaoYang Huang**, School of Civil and Ocean Engineering, Jiangsu Ocean University, Lianyungang 222005, Jiangsu, China, e-mail: 2022120813@jou.edu.cn
*** Corresponding author: Muhammad Nasir Amin,** Department of Civil and Environmental Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia, e-mail: mgadir@kfu.edu.sa
*** Corresponding author: Suleman Ayub Khan,** Department of Civil Engineering, COMSATS University Islamabad, Abbottabad 22060, Pakistan, e-mail: sulemanayub@cuiatd.edu.pk
Kaffayatullah Khan, **Muhammad Tahir Qadir:** Department of Civil and Environmental Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia

1 Introduction

Concrete is a composite element characterized by its heterogeneity, multi-scale nature, and multi-layered composition, collectively contributing to its intricate structure. The macro-physical and mechanical behaviors of the material exhibit characteristics such as inconsistency, uncertainty, and non-linearity, which indicate its complex morphology. Although there has been extensive research on the macro and micro characteristics of cement-based substances, there has been comparatively less emphasis on comprehending the microstructure of concrete. The arrangement of pores and distribution of pore sizes in concrete are critical factors that directly impact its durability and strength [1]. Hence, developing a rapid assessment technique capable of evaluating these characteristics of concrete would hold significant value in forecasting the efficacy of concrete structures. The correlation between concrete's porosity and transport characteristics has garnered excessive attention in academia. These characteristics include the chloride ion migration, diffusion coefficients of gas, electrical resistivity, and gas or water permeability [2–4]. Figure 1 illustrates the empirical correlation of the compressive strength (CS), porosity, and permeability of concrete. In general, the permeability of cement composite tends to increase with an increase in porosity, whereas its mechanical strength tends to decrease. Hence, the criterion of porosity plays a fundamental role in assessing the longevity and functionality of concrete structures subjected to harsh environmental conditions [5].

Concrete can be classified as a chemically bound ceramic product when formulated with a hydraulic binder. The

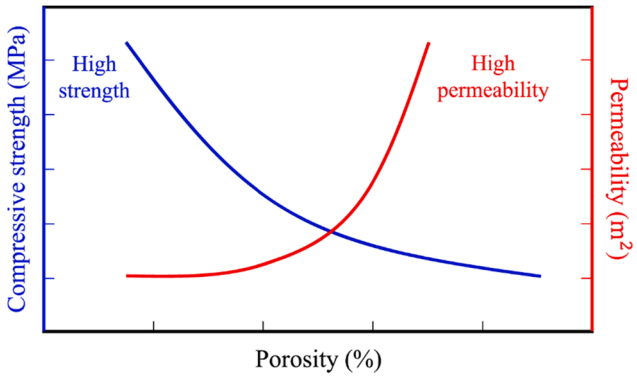

Figure 1: Relationship between porosity, permeability, and CS [6].

cement reaction with water leads to the formation of a composite material comprising a solid phase and a network of pores. Pores are an inherent characteristic of concrete. The pore system plays a pivotal role in determining the critical characteristics of concrete, precisely its strength [7]. Insufficient compaction can also lead to the development of pores inside concrete. The pore structure seen in the mortar or concrete exhibits a distinct arrangement from the pores observed in well-compacted mortar or concrete manufactured separately with identical amounts of the corresponding constituents. The disparity in both pore systems can be attributed to interfacial transition zone (ITZ) pores at the interface between mortar and aggregate [8]. The porosity of cured cementitious composite is contingent on the water-to-binder (w/b) ratio. The w/b likewise regulates the porosity of ITZ in concrete. The increase in w/b leads to an expansion in the dimensions of the pore spaces within the hydrated cementitious composite. The reduction in porosity occurs as the duration of curing extends and the hydration process improves, primarily due to the sealing or connection of large-dimension voids by calcium-silicate-hydrate (CSH) hydrogel pores. A standard model for determining the volumetric makeup of the hardened cementitious composites has been established based on the w/b and the extent of cement hydration [9]. According to earlier experimental findings, it has been observed that an augmentation in both the size and percentage of coarse particles results in a corresponding increase in the porosity at the ITZ. Consequently, this leads to a decrease in the durability of conventional concrete [10].

The weight ratio between coarse and fine aggregates (CA/FA) significantly affects concrete's permeability, porosity, and tortuosity [11]. Supplementary cementitious materials (SCMs) are currently employed as a means of partially substituting Portland cement (PC) to augment the durability as well as the strength of concrete [12,13]. Furthermore, SCMs, for instance, fly ash, which is a result of the burning of coal in the thermoelectrical factory [14], or ground granulated blast furnace slag (GGBFS), which is a derivative of pig iron manufacturing, have the potential to facilitate cleaner production processes by effectively mitigating CO_2 emissions [15]. The primary benefit of GGBFS in concrete is ascribed to its inactive hydraulic reaction. This reaction plays an imperative role in enhancing the hydration of cement by compacting the concrete's compound and improving the pore formation. This phenomenon leads to decreased permeability and a surge in CS for concrete that undergoes maturation over extended periods. The alteration in the mineral constituents of the cement has the potential to improve the concrete's ability to capture chloride ions and elevate its electrical resistance [16]. The decline in permeability of concrete modified with fly ash can be ascribed to the mutual impact of the diminished quantity of water necessary for specific workability and a more processed void structure resulting from the pozzolanic reaction. The benefits of the pozzolanic process become increasingly apparent in well-cured concrete due to its extended lifespan. Hence, the curing state, whether water or air curing, is an additional significant component that influences the porosity of concrete [17]. Moreover, the use of superplasticizers (SPs) has the potential to significantly decrease the extent of mix-up water, hence promoting the development of a more compact pore arrangement [18].

Due to the diverse compositions of cementitious compounds and the intricate reaction of cement hydration over time, the evolution of the pore system in concrete exhibits a high level of complexity, making it challenging to represent accurately by analytical modeling. The uncertainty of the pozzolanic reactivities of GGBFS and fly ash poses a significant difficulty. The chemical structure of SCMs exhibits considerable variability, posing difficulties in accurately determining the reactive fraction of these materials. Papadakis [19,20] introduced a theoretical framework for predicting fly ash-based cementitious compounds' chemical makeup and related volume. The prototype deliberates the stoichiometry of PC's hydration process and the pozzolanic reactivity of fly ash while also considering the molar weight of the components and byproducts involved. Nevertheless, this prototype operates under the assumption of complete hydration of PC and the occurrence of pozzolanic reactivity involving fly ash. Consequently, the prototype cannot account for the time-reliant changes in porosity. Similarly, Salih *et al.* [21] used distant modeling techniques, including artificial neural network (ANN), M5P tree, and nonlinear and linear regression (LR), to estimate the CS of fly ash-based modified concrete. Mohammed *et al.* [22] used 379 data points to compute the CS of fly ash-based modified concrete using neuro-imperialism and neuro-swarm

models. Piro *et al.* [23] studied the optimization of full quadratic (FQ), multi-logistic regression (MLR), and ANN to predict the electrical resistivity and CS of concrete modified with GGBFS and steel slag as aggregate replacement. Piro *et al.* [24] developed MLR, ANN, and FQ models to forecast the CS and electric resistivity of concrete modified with steel slag as an aggregate replacement.

Several investigations have been conducted to compute concrete porosity [25–27]. However, manufacturing concrete compounds to compute the porosity and pore structure requires a significant allocation of resources, money, time, and labor. The process involves methodically choosing substances and their corresponding amounts, guided by a series of thorough experiments. Various scholars have proposed statistical methods to predict the advancement of cement hydration levels [28,29]. An empirical cement hydration model estimates the time-reliant changes in porosity and chloride diffusion coefficient in concrete [30]. To formulate empirical predictions regarding the permeability characteristics of concrete, Khan [31] employed multivariate regression analysis. This investigation entailed the computation of porosity by considering different constituents of the concrete mixture, such as the percentage of fly ash, the ratio of micro-silica, and w/b.

While these models were successfully employed to predict the porosity of concrete, their accuracy heavily relies on the underlying assumptions. Applying machine learning (ML) methods can provide numerous advantages in solving challenges and enhancing the efficiency of computing concrete porosity. In contemporary times, ML algorithms have been effectively utilized across several academic fields to predict multiple attributes precisely. Similarly, the implementation of these solutions in civil engineering has the capacity to generate significant advantages in terms of optimizing the testing process and thereby lowering time consumption. The implementation of such strategies has been utilized. Several methodologies have been utilized in the calculation of the strength and durability features of cementitious composites, such as genetic programming (GP) [32], ANN [33–35], decision tree (DT) [36], and support vector machine (SVM) [37]. The model findings demonstrate a significant correspondence between the anticipated values and the empirically derived concrete characteristics indicating that ML holds promise as a tool for modeling concrete with intricate mixed compositions. Nevertheless, there is a scarcity of research focusing on utilizing ML techniques for predicting concrete porosity. However, limited studies are available to describe the utilization of ML algorithms to forecast the porosity of concrete. For instance, using the ANN model, Pereira *et al.* [38] studied porous concrete's airflow resistivity, open porosity sound absorption, and tortuosity. Similarly, the study of Sathiparan *et al.* [39] demonstrated the utilization of various ML algorithms like ANN, SVM, LR, RF, *k*-nearest neighbor, and DT to predict the permeability and porosity of concrete. Cao [6] developed three models, including XGBoost, RF, and XGB, to forecast the porosity of high-performance concrete. Wu *et al.* [40] studied the predictive performance of the ANN model using a permeability dataset of 3,252 observations. Similarly, the ML methodologies employed in the literature to compute the porosity of concrete are provided in Table 1.

The main focus of this investigation is to assess the microstructural characteristics of concrete, with a particular emphasis on porosity. To achieve the objective, ML techniques rooted in artificial intelligence were utilized, including adaptive boosting (AdaBoost), LR, SVM, and DT. The efficacy and precision of these algorithmic models exhibit a distinct correlation with the quantity of data elements. A comprehensive database of 240 data components regarding concrete porosity was compiled utilizing experimental results from previous studies. The modeling technique considered seven input elements: the quantity of binder, proportion of fly ash, water/binder ratio, percentage of slag, ratio of coarse-to-fine aggregates, SP, and number of curing days. A comparative evaluation has examined the outcomes derived from the ML models,

Table 1: Previous studies conducted to predict the porosity using ML methodologies

Ref.	Methods employed	Properties studied	Results
[41]	GP with the combination of ANN	Tortuosity, CS, porosity	The model showed excellent results with a high correlation value
[6]	XGB, RF, GBT	Porosity	GBT outperformed
[42]	XGB, RF, SVM	CS, porosity	XGB outperformed
[43]	LR, AdaBoost, RF, Bagging, XGB, Cat-Boost, Light-GB, DT	CS, flexural strength, porosity, workability	Models showed excellent results with a high correlation value
[44]	GBT, RF	Porosity	GBT outperformed

Annotation: XGBoost (XGB), Random Forest (RF), Gradient Boosting Tree (GBT), Support Vector Machine (SVM), Linear Regression (LR).

specifically SVM, AdaBoost, DT, and LR. Statistical methodologies were utilized to compute and compare the model's predictions, ability to generalize, inaccuracies, and proficiency. Furthermore, this study implemented the RReliefF algorithm to determine the key input variables that have the most significant influence on the porosity of concrete.

2 Data description

The rationale of this study was to predict the porosity of concrete. ML approaches necessitate the utilization of several inputs to produce the desired target variable. The data employed in this investigation for predicting the porosity of concrete are derived from previously published literature with specific standards for selection [45–50]. Initially, the concrete was blended with PC, which could be partially substituted with either GGBFS or fly ash. Furthermore, it is worth noting that the concrete matrix comprises both coarse and fine particles. The influence of carbonation on voids arrangement alteration was not considered in the concrete sample. In addition, a well-maintained equilibrium was preserved for each concrete variant, namely PC concrete, fly ash-based concrete and GGFBS-based concrete. The experimental setup for evaluating the porosity of concrete is shown in Figure 2(a)–(d). Seven distinct properties were selected as inputs to estimate the porosity of concrete. The specific information regarding these elements is listed in Table 2. The most critical stage while computing concrete properties *via* ML is the compilation of a comprehensive dataset. The accurate predictability of ML models is highly dependent on the quality and quantity of the dataset. As per literature studies, the adequate performance of the model is contingent upon the quantity of data in relation to inputs. An optimized model necessitates a ratio greater than 5 to analyze the relationship between the necessary variables

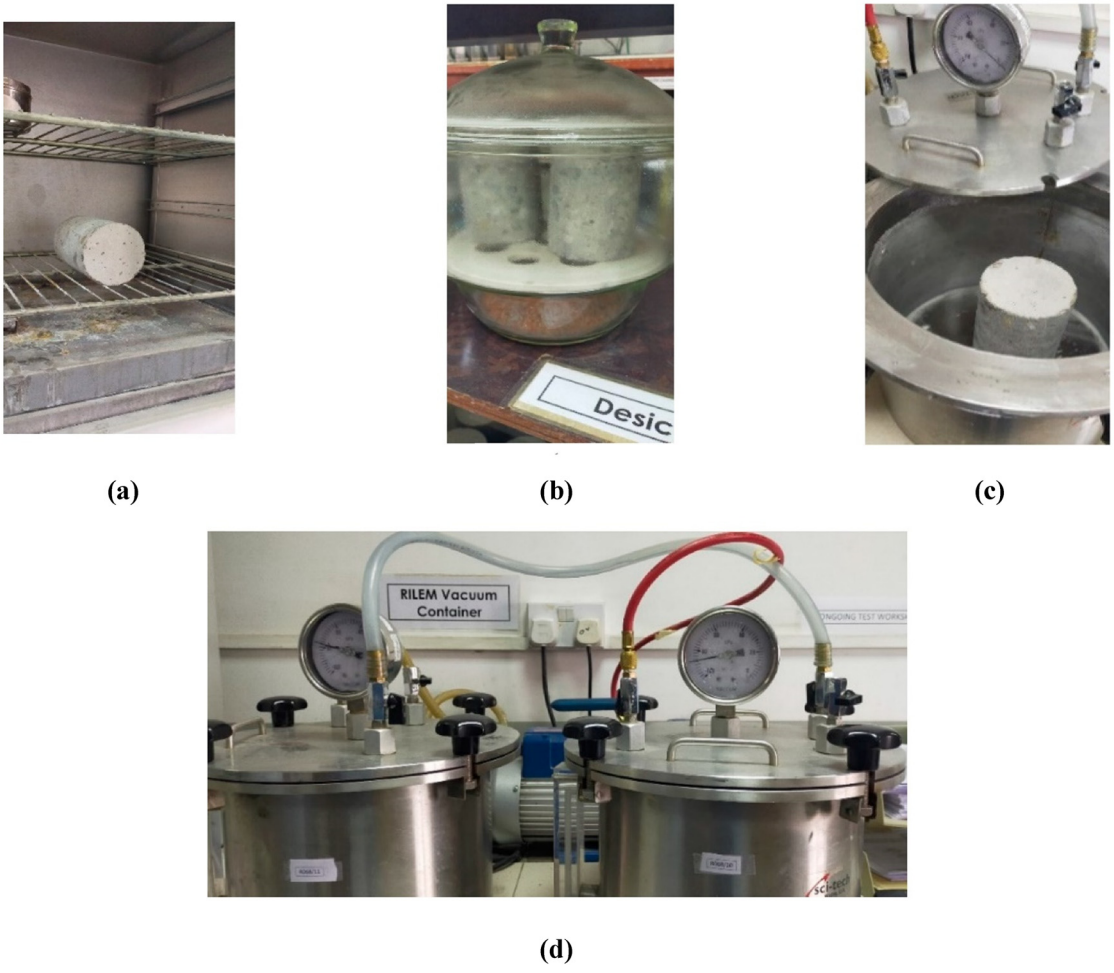

Figure 2: Experimental setup for computing porosity of concrete: (a) oven drying of the sample; (b) sample cooling in a desiccator; (c) sample inside a vacuum container before beginning the test; and (d) sample in vacuum conditioning [53].

Table 2: Specification of input features

Variables	Abbreviations	Unit
Binder	—	$kg \cdot m^{-3}$
Coarse aggregates/fine aggregates	CA/FA	—
Water/binder	w/b	—
Fly ash	—	%
Curing time	CT	Days
Ground granulated blast furnace slag	GGBFS	%
Superplasticizer	SP	%

effectively. The current study utilized a dataset of 240 records to calculate the porosity of concrete with seven potential inputs, hence fulfilling the requirement for the adequate performance of ML models [51,52]. A total of 240 data points were used to train the ML models in which 75 observations were of standard PC concrete, 120 observations were of fly ash-based modified concrete, and 45 records were of GGBFS-based modified concrete. Figure 3 shows the frequency distribution histograms of all the datasets utilized in constructing models.

The process of normalizing and standardizing data is often considered crucial in preparing data for training ML models. Normalization may not be necessary or appropriate when the data are categorical or sparse. Alternative preparation techniques like data preprocessing may better suit specific situations [55]. The present study employed data preparation to assess its appropriateness for ML modeling. The primary processing of gathered data is a crucial and central phase in advancing and improving ML models. Data preparation encompasses a variety of commonly employed processes, including the management of missing data, the encryption of variables, the identification and treatment of outliers, and the partitioning of the data into sections. The database utilized in the present study was devoid of any missing information or outliers. Although the research did not employ a multivariate strategy to find outliers, a reliable data pre-treatment procedure successfully guaranteed the non-existence of abnormalities. Before training the ML models, a methodology was employed to detect outliers for each variable. The procedure involved the detection and exclusion of any data, if applicable that deviated from the predefined acceptable ranges. Additionally, a thorough demographical and statistical assessment was conducted on the dataset to sense and rectify any potential outliers. The statistical measurements offer boundaries, both higher and lower. The study incorporates various statistical metrics, including median, skewness, standard deviation, mode, standard error, mean, and kurtosis. The median, as well as the mode score of the SPs data, was found to be "0" due to the prevalence of values that are either 0 or in extreme proximity to 0. The frequency of values in proximity to 0 within the dataset for SPs can be credited to the constrained quantity of these constituents that were blended into the mixture. Skewness is the form of statistical indicator employed to quantify the severity of imbalance in the frequency distribution of a particular attribute with continuous values corresponding to its mean. In the present context, negative results typically indicate the presence of a long tail on the left side of the bell curve. Kurtosis is a statistical indicator used to evaluate the extent of the tail's lightness or heaviness in a given data, as well as its suitability for a particular normal distribution. This provides valuable comprehension of probability distribution throughout the vertical axis. Table 3 provides an in-depth review of the statistical summary performed on all independent data features.

The evaluation of the relationship between inputs has been recognized as a crucial measure for assessing their influence on the target inside the employed dataset. Figure 4 depicts a plot of the Pearson correlation coefficient, which showcases the statistical relationship between variables for the complete dataset employed in the present study. Correlation coefficients are utilized as a quantitative matrix to evaluate the intensity and direction of the linear correlation between two entities. The correlation tool is a frequently employed gadget in the domain of statistics to examine the interaction between different attributes. Potential outcomes encompass an entirely (−ive) correlation, symbolized by −1, a wholly (+ive) correlation, expressed by +1, and the non-appearance of any correlation, indicated by 0. A +1 score implies that the increase in one entity is consistently associated with an increase in another entity. At the same time, a −1 correlation says that a decline in another entity usually accompanies the increase in one entity. The input variable that had the peak level of prominence was fly ash, which revealed a more pronounced (+ive) correlation with the target parameter (porosity). The association between these variables was additionally substantiated by a correlation value of 0.32. A positive connection was seen between the CA/FA, fly ash, and w/b, while the correlation between the binder, GGBFS, SP, and CT showed a (−ive) alliance with the target variable. The complete database was arbitrarily subdivided into two definite subsets: training data, comprising 70% of the total data points, and testing data, comprising 30% of the remaining data.

3 Research methodology

ML algorithms are utilized in several research fields to anticipate materials' serviceability. As nonlinear interaction exists between various concrete components, which

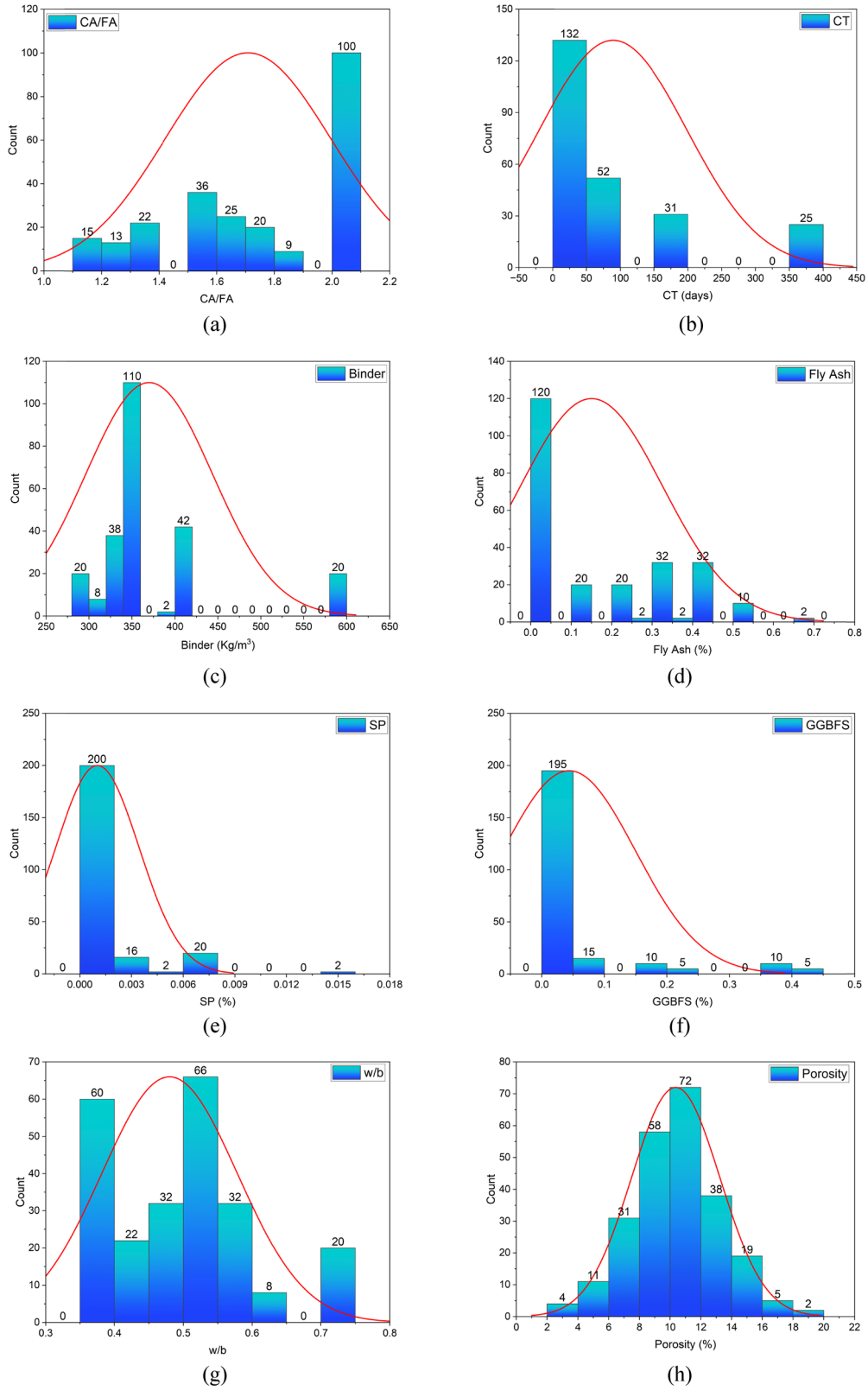

Figure 3: Frequency distribution histograms: (a) coarse aggregate/fine aggregate, (b) curing time, (c) binder, (d) fly ash, (e) SP, (f) slag, (g) water/binder, and (h) porosity (parameters similar to the study of Tian *et al.* [54]).

Table 3: Statistical explanation of inputs and target (parameters similar to the study of Tian *et al.* [54])

	w/b	Binder	Fly ash	GGBFS	SP	CA/FA	CT	Porosity
Mean	0.480	369.942	0.150	0.044	0.001	1.708	89.358	10.362
Sample variance	0.010	5469.394	0.031	0.011	0.000	0.082	11904.984	8.293
Standard error	0.006	4.774	0.011	0.007	0.000	0.019	7.043	0.186
Median	0.500	350.000	0.050	0.000	0.000	1.722	28.000	10.325
Mode	0.500	350.000	0.000	0.000	0.000	2.000	28.000	10.300
Standard deviation	0.098	73.955	0.176	0.106	0.002	0.287	109.110	2.880
Maximum	0.700	591.000	0.670	0.400	0.016	2.000	365.000	18.047
Kurtosis	−0.144	3.965	−0.795	5.485	11.909	−1.289	1.511	0.267
Range	0.350	296.000	0.670	0.400	0.016	0.806	364.000	15.647
Skewness	0.550	2.117	0.721	2.552	3.143	−0.376	1.610	−0.040
Minimum	0.350	295.000	0.000	0.000	0.000	1.194	1.000	2.400
Sum	115.203	88786.000	36.027	10.479	0.244	410.009	21446.000	2486.797

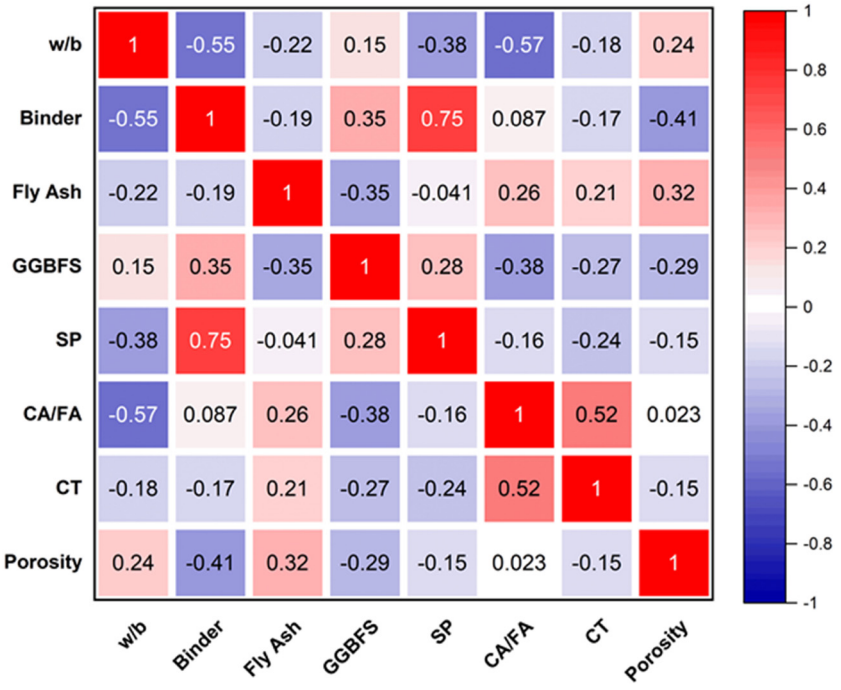

Figure 4: Pearson correlation coefficient chart.

includes binder, coarse aggregates/fine aggregates, water/binder, fly ash, curing time, GGBFS, and SP, this research utilizes AdaBoost, LR, DT, and SVM techniques to predict the porosity of cement-containing compounds. These techniques are well recognized for capturing the non-linear pattern between the concrete components, ensuring the problem of oversimplification and overfitting. These models utilized sufficient parameters and hyperparameters, which allowed us to learn the complex pattern between inputs and output variables, enabling the algorithm to predict the outcomes precisely. These approaches are adopted due to their extensive practice, reliable predicting capacities in related research, and identification as the most efficient data-driven techniques. The utilization of the correlation coefficient (R^2), which falls within an interval of 0 to 0.99, is a prevalent method for evaluating the level of accuracy in foreseeing properties by contrasting them to their observed values. A higher R^2 signifies that the selected algorithm has produced satisfactory results. The in-depth flowchart of the current investigation is depicted in Figure 5. The model's algorithms are run in Orange (3.35.0), a software platform based on Python. The inclusive arrangement of models in orange software is provided in Figure 6.

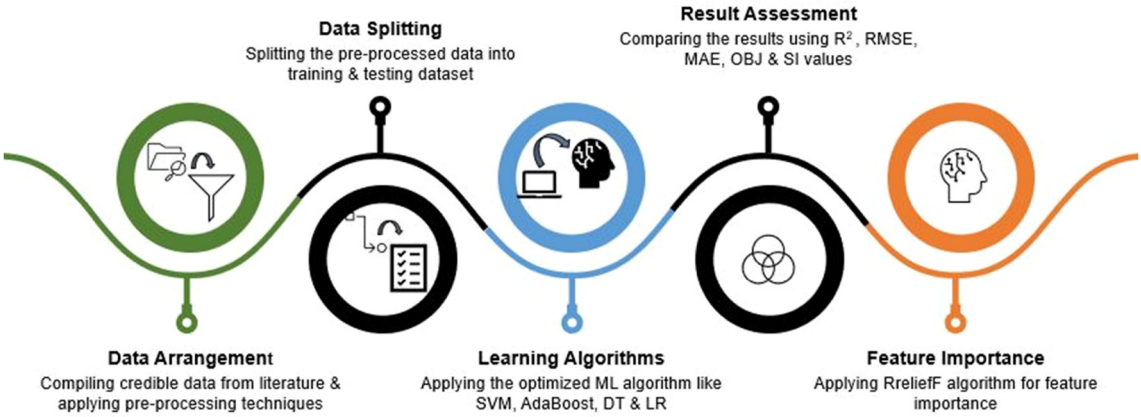

Figure 5: Flow diagram of research.

3.1 DT algorithm

A DT is a controlled learning technique that may handle input and target variables that are either categorical or continuous. The algorithm is capable of making predictions for both categorical data, using a classification tree, and continuous variables, using a regression tree. The algorithm operates by iteratively dividing a dataset into smaller segments using the most influential attribute, defined by a splitting criterion such as variance reduction, information gain, or gini impurity. Each division yields nodes, and this process continues until a stopping criterion is met. Typically, the threshold is reached when each of the points in a node is part of the same class (in classification) or when further divisions do not appreciably enhance predictions [56]. The DT is an extensively employed regression approach that offers an understandable framework for analysis. The key benefit of DT is its ability to imitate decision-making in humans, rendering it more accessible compared to other ML algorithms. A rudimentary hierarchical structure is constructed to model prospective outcomes, consisting of root systems, limbs, and leaf nodes, representing the predictions [57]. The outcomes, as shown by the leaf nodes, are situated at the terminus of the flow chart representing the DT [58]. The flow chart commences with root buds and thereafter turns to branches. The hierarchical structure of the given database commences with the primary root bud, which commonly functions as a symbolic description of the

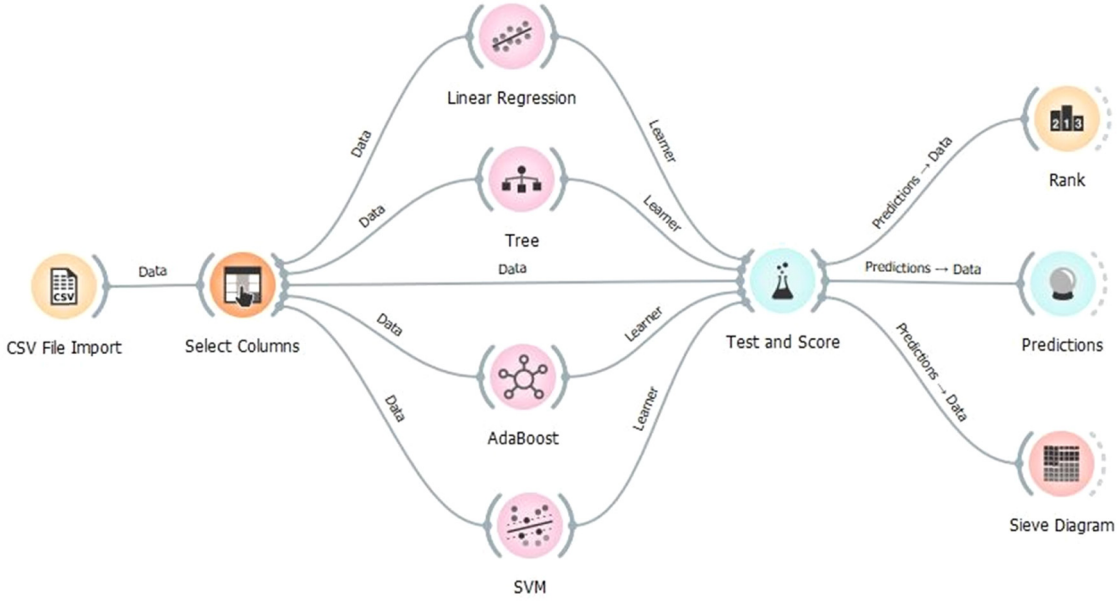

Figure 6: Model arrangements in software.

complete database. The effectiveness of a particular modeling approach is determined by the functions assigned to each root bud. The classification concept is determined by the direction every parent node (Root) takes as it advances the leaf, as shown in Figure 7. These buds are categorized into three distinct geometric shapes: triangles, rectangles, and circles. The DT classification technique is widely considered to be bare, with a simplicity that facilitates comprehension and use [59]. The functions applied to develop the DT algorithm for the prediction of porosity are summarized in Table 4.

3.2 LR algorithm

LR is a statistical framework for determining the linear connection between a single response variable (referred to as a dependent variable) and one or more explanatory factors (referred to as independent variables). The fundamental concept is to identify a linear correlation that most accurately corresponds to the given data. The multiple LR algorithm is an expanded version of the standard regression algorithm that aims to establish the correlation between a numerical target parameter and two or more independent factors [61]. The proposed model postulates that a linear function of the experimental values of the independent variables can adequately represent the fitted probability of the occurrence. The computational problem that LR tackles involves the process of fitting a hyperplane onto the nth-dimensional space, where n corresponds to the number of independent variables. In the context of a structure with nth independent variables, denoted as x's, and a single output parameter, denoted as Y, the overarching objective of the generic least-squares problem is to ascertain the unidentified variables of the LR model. This study aimed to assess the suitability of LR due to its simplicity. Eq. (1) presents the general expression for LR models [62]:

$$y = \beta x + \varepsilon, \tag{1}$$

where y is the target variable, β is the regression coefficient, x is the input features, and ε is the error term.

This research utilizes a specific methodology to efficiently approximate several linear equations illustrating the correlation between the porosity and the provided independent factors. To boost the prediction capabilities of the LR algorithm, polynomial parameters are created by applying multiple degrees of polynomials to the fundamental features.

Table 4: Functions adopted for the DT algorithm

Parameters	Assigned function
Least no. of instances in leaves	2
Maximum tree depth	50
Induce binary tree	Yes
Minimal subset	5
Stopping criteria	95%

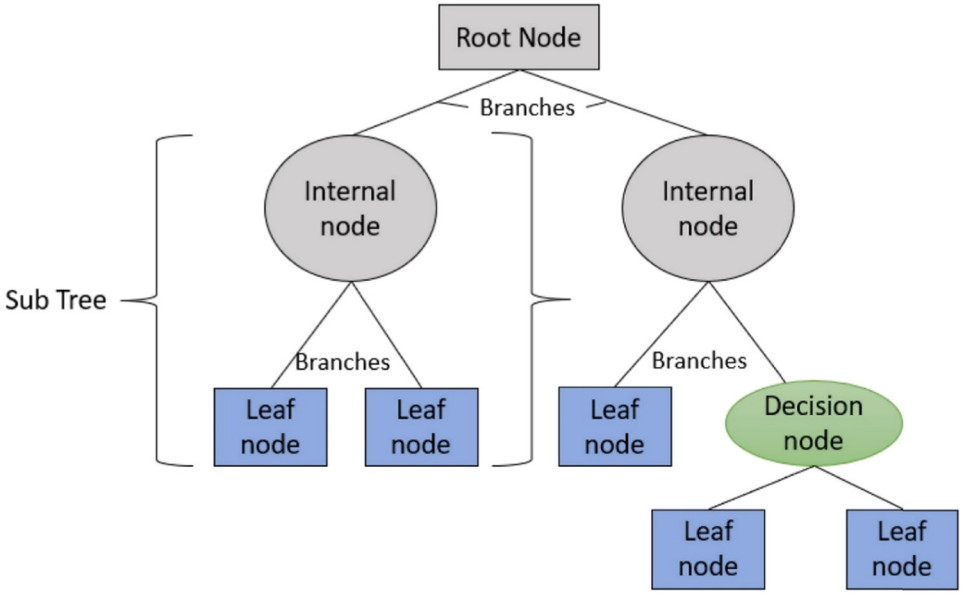

Figure 7: Hierarchical structure of DT [60].

3.3 SVM algorithm

The SVM, introduced by Vapnik [63], is comprehensively utilized in several fields, such as regression, classification, and predictions. The SVM utilizes a kernel function to translate the independent dataset into a higher-dimensional space, enabling it to address non-LR issues [47–49] effectively. The *k*-function treats data points falling within a predefined tolerance level of the true value as essentially the same. The actual distance is equivalent to the spatial separation relating to the two hyperplanes. Minor divergences are not subject to strict monitoring. However, significant deviations are not accepted under any circumstances. The essential factors that ensure the accuracy and generalization capability of the SVM model are sample data processing, parameter optimization choices, and *k*-function. The SVM algorithm can be categorized as either a binary classifier, in which the targeted variable adopts just two values (0 and 1), or a regression classifier, in which the targeted variable takes on indefinite fundamental values. The SVMs as regression classifiers are commonly employed for the construction of the input–output model because of their successful resolution of non-LR challenges [64].

The primary step in SVM as a regression classifier is mapping the input to an *n*th-dimensional parameter space utilizing a predetermined mapping technique. Non-linear kernel functions are utilized to effectively model the high-dimensional parameter space, resulting in enhanced separability of independent datasets when differentiated from their equivalents in the primary independent space. The linear classifier in the space is denoted as *f(a, b)*, and it can be scientifically denoted by the following equation [62]:

$$f(a, b) = \sum_{i=1}^{n} b_i g_i(x) + c, \tag{2}$$

where b_i is the weight vector, which can be computed by reducing the normalized risk function, which incorporates the empirical risk, $g_i(x)$ is the collection of non-linear transformations that operate on the independent space, and c is the bias value.

The mathematical equation of the kernel function is given in Eq. (3) [62]:

$$K(x, x_j) = \sum_{j=1}^{n} g_j(x) g_j(x_j). \tag{3}$$

SVM's *k*-functions, for instance, linear, polynomial, sigmoid, or radial basis functions, are adopted to locate the support vectors, which are the critical data points that determine the position of the decision function during training. The default *k*-functions are contingent upon the specific *k*-class and the employed software. The parameters selected to optimize the SVM model for forecasting the porosity of concrete are summarized in Table 5.

3.4 Adaptive boosting algorithm

A single DT, when used as a separate model, is sometimes referred to as a poor learner due to its inherent limitations in terms of predictive power and generalization ability. The likelihood of attaining a robust learner by amalgamating several weaker learners is an active research area. The central concept is to prioritize instances that pose the most significant challenge in terms of classification. AdaBoost operates in an iterative manner: with each round, it modifies the weights of wrongly classified instances, increasing their importance in the subsequent training phase. During each step of this ongoing process, a new weak learner is introduced and trained to rectify the errors made by the prior ensemble. The speculation was proven by Freund *et al.* [65] in 1990, establishing the fundamental principles for the boosting algorithm, a technique that consecutively combines numerous weak learners. As demonstrated in Eq. (4) [66], incorporating a new tree algorithm into the overall structure eliminates the typical tree, with just the most robust tree included. Through the process of iterative computation, the performance of the entire model will continuously enhance. Following the acquisition of the primary rudimentary tree model, specific sections within the dataset are accurately identified, while the others are erroneously labeled:

$$f_j(x) = f_{j-1}(x) + \operatorname{argmin}_k \sum_{i=1}^{n} a y_i f_{j-1} x_i + k x_i, \tag{4}$$

where $f_j(x)$ represents the entire model, $f_{j-1}(x)$ represents the entire mode in the prior round, y_i represents the

Table 5: SVM model generalization function

Parameters	Assigned function
Type	V-SVM
	Complexity bound (v) = 0.40
	Regression cost (c) = 1.00
Kernel function	Polynomial
	equation = $(g(x)y + c)^d$
	g = 0.10
	c = 0.90
	d = 4.00
Optimization	Tolerance = 0.0001
	Iteration limit = 500

forecasted outcome of the ith tree, and kx_i represents the freshly incorporated tree.

The AdaBoost algorithm is an iterative method for enhancing the performance of weak classifiers by iteratively learning from the data. This iterative process aims to increase the algorithm's overall classification ability. The first poor classifier is derived by the process of training on the provided samples, wherein the misclassified samples are mixed alongside the untrained data to create a novel trained prototype. Moreover, the succeeding poor classifier is acquired through the process of learning from this prototype. The incorrect prototype is merged alongside the untrained data to create a novel trained prototype, which may be used to derive the third poor classifier for training purposes. Afterward, iteratively executing this procedure multiple times can eventually obtain the reboot-resilient classifier. To boost the accuracy of classification, the AdaBoost methodology assigns multiple weights to the data [65]. The appropriately categorized samples are assigned relatively low weights, while the misclassified samples need to be given higher weights. This compels the algorithm to allocate more importance to the misclassified data [67]. Figure 8 provides a comprehensive depiction of the computational procedure employed by the AdaBoost method. To train any basic DT model, altering the weight distribution that appears for each sample inside the dataset is mandatory. Given that each training data point changes, the training results will similarly demonstrate variability. Consequently, the cumulative sum of all outcomes is obtained [68]. Table 6 provides detailed insight into all the parameters adopted to run the AdaBoost algorithm.

Table 6: Functions utilized to run the AdaBoost model

Parameters	Assigned value/function
Basic estimator	Tree
Classification algorithm	SAMME.R
Learning rate	0.50
Regression loss function	Linear
Estimator's number	100

3.5 Model validation and evaluation criteria

Model validation refers to assessing the extent to which a model accurately represents the real world, considering its intended purpose. Engineers commonly employ qualitative validation methods, such as graphical comparison, to assess the accuracy of model predictions concerning experimental data. Nevertheless, it is necessary to employ statistics-based quantitative methods to complement subjective judgments and methodically consider error and unpredictability in modeling forecasting and experimental observation [70]. Most of the previous research focuses on evaluating the model's performance and validation using statistical matrices [33,51,55,71]. As this technique accurately foresees the generalization capabilities of ML models, this study also utilized statistical error analysis to validate the model performance. The evaluation of the constructed ML models involved the utilization of statistical metrics, such as correlation coefficient (R^2), objective function (OBJ), mean absolute percentage error (MAPE), scatter index (SI), root mean square error (RMSE), and mean absolute error (MAE). The R^2 factor for predicted outcomes serves as a measure of the accuracy of the utilized algorithms. The R^2 factor is used to quantify the disparity between the projected algorithms and the target metrics [72]. A numerical value closer to "0" implies a greater level of deviation, whereas a numerical score closer to "1" suggests a lesser level of deviation. The reduced mistakes perceived in the statistical score indicate the enhanced precision of the created algorithms. The statistical assessment of the accuracy of ML algorithms was conducted by employing Eqs. (5)–(9), which were sourced from the study of Cao *et al.* [73]:

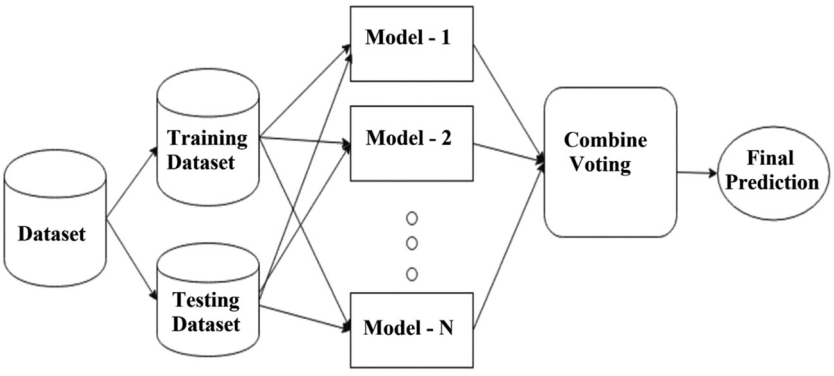

Figure 8: Schematic diagram for AdaBoost algorithm [69].

$$\text{RMSE} = \sqrt{\sum \frac{(P_i - E_i)^2}{n}}, \tag{5}$$

$$\text{SI} = \frac{\text{RMSE}}{y'}, \tag{6}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |P_i - E_i|, \tag{7}$$

$$\text{OBJ} = \left(\frac{\text{MAE} + \text{RMSE}}{R^2 + 1}\right), \tag{8}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n} \frac{|P_i - E_i|}{E_i}, \tag{9}$$

where n is the number of datasets, P_i is the predicted outcomes, E_i is the experimental values, and y' is the average of predicted outcomes.

3.6 RReliefF algorithm

The Relief algorithm selects features as reported by their relevance for the output function [74]. The Relief analysis's core concept resembles the fundamental principles underlying the *k*-nearest neighbor method. Suppose a particular characteristic is regarded as valuable. In that case, it is anticipated that the nearest proximity of entries attributed to the same class will be adjacent to the range specified for that attribute in contrast to the nearest proximity of entries from all other classes. The score determined using the relief method is given by Eq. (10) [75]. The feature space that is taken into consideration by the relief method is denoted as "*ij*," which represents the number of entries used for computing the close-hit (CH) and close-miss (CM) values. Feature X is a factor that is taken into consideration while calculating the relief score:

$$w = \frac{w - \text{diff}(x_{ij}, \text{closehit}_{ij})^2 + \text{diff}(x_{ij}, \text{closemiss}_{ij})^2}{m}. \tag{10}$$

As an illustration, when considering feature X, the algorithm computes the CH metric within the confines of the same class. The term "CH" refers to the minimum distance across features within the sample space encompassed by the same class. On the other hand, "CM" refers to the distance across feature X from the sample space encompassed by different classes. The original relief metric is designed for binary classification tasks, while ReliefF is an expansion that can handle multinominal classification tasks [76]. The modification of ReliefF for the purpose of regression is referred to as RReliefF [77]. The RReliefF algorithm is designed to handle data characterized by noise, incompleteness, and several classes. The RReliefF method for multinominal classification employs *K*-nearest hit and *K*-nearest miss accumulation techniques for *K*-classes. The RReliefF count provides a comprehensive perspective on the variable quality judgment through estimation. The RReliefF process considers a specified number of samples to produce attribute scores, comprehensively assessing variable quality globally. The chain of Relief algorithms has gained significant recognition for their exceptional performance. They have been widely applied in various domains, such as signal identification, fault diagnosis, and computational imagery processing and segmentation. Scholars have developed methodologies to examine the process of parameter selection using hierarchical learning within a specific subset of parameters (or a defined subspace) to determine the most optimal subspace [33].

4 Results and analysis

4.1 LR outcomes

The analysis focused on the outcomes of the prediction and the error associated with the LR algorithm. Figure 9(a) exhibits the association between the experimental and anticipated outcomes, as indicated by an R^2 score of 0.763. This correlation implies that the precision of outcomes is inferior to that of the DT model. Figure 9(b) demonstrates the distribution of observed, forecasted, and errors in the LR model. The training set exhibited maximal upper limit, minimal limit, and mean limit error readings of 4.52, 0.00, and 1.13%, respectively. Overall, 51.67% of the errors were found to be below 1%, while 45.00% were within the range of 1–3%. The remaining 3.3% of error values exceeded 3%.

4.2 DT outcomes

The values derived from the DT approach exhibit a satisfactory level of accuracy and only a minimal magnitude of divergence between observed and predicted outcomes. Figure 10(a) presents the data analysis evaluation of the observed and projected results for the porosity of cementitious composites. R^2 indicates that the model accurately predicts outcomes with a value of 0.838. The DT model outperformed the LR model, as suggested by a higher R^2 score. DT models can capture complex interactions and non-linear patterns within the data. This characteristic makes them particularly well-suited for tasks in which the linear assumptions of LR may not be valid. Figure 10(b) displays the distribution of data obtained from experiments, projected scores, and errors for the DT algorithm. The training set was

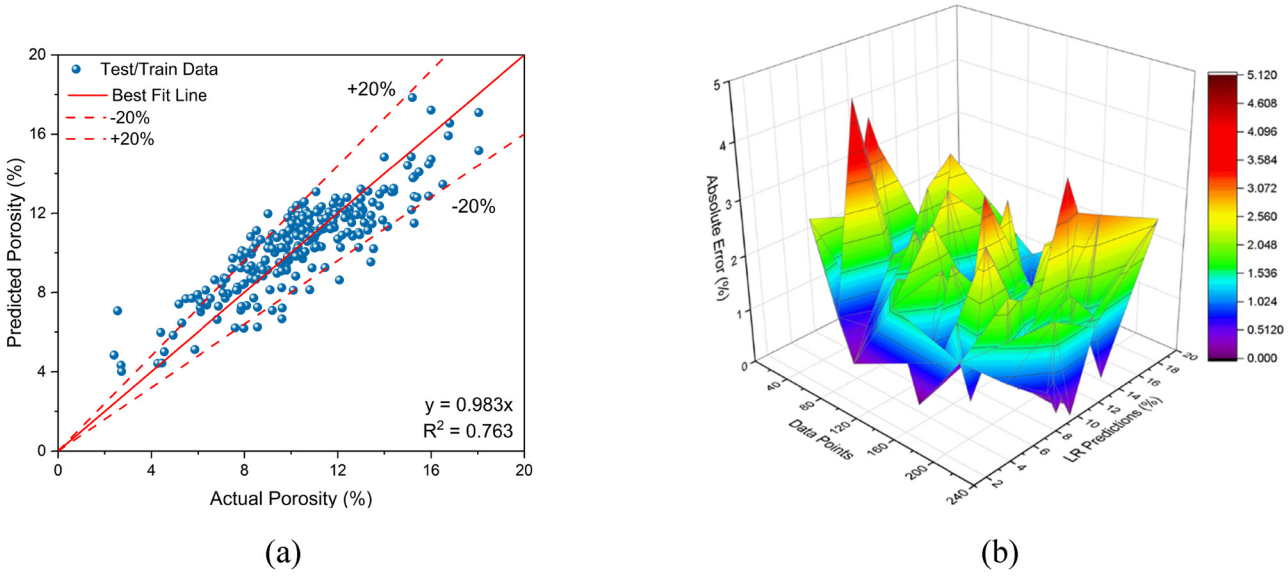

Figure 9: LR model outcomes: (a) R^2 graph and (b) error tracing.

analyzed to find the maximum upper limit, lowest limit, and mean limit values, which were found to be 5.55, 0.00, and 0.82%, respectively. Furthermore, approximately 73.75% of the error values observed in the study were found to be less than 1%. Around 23.75% of the error values also fell within the 1–3% range, while approximately 2.50% exceeded 3%.

4.3 SVM outcomes

The analysis focused on the prediction outcomes and errors associated with the SVM model. Figure 11(a) showcases the interaction between experimental and anticipated results, indicating higher precision in outcomes relative to the DT and LR models, as supported by an R^2 value of 0.870. SVMs have exceptional performance in identifying ideal hyperplanes or non-linear transformations that optimize the margin between distinct classes, hence offering a robust separation between data points. Moreover, SVM models have a lower susceptibility to overfitting in comparison to DT and LR algorithms and possess the ability to effectively handle datasets with a high number of dimensions. Figure 11(b) demonstrates the distribution of observed, forecasted, and errors in the SVM model. The training set exhibited maximal limit, minimal

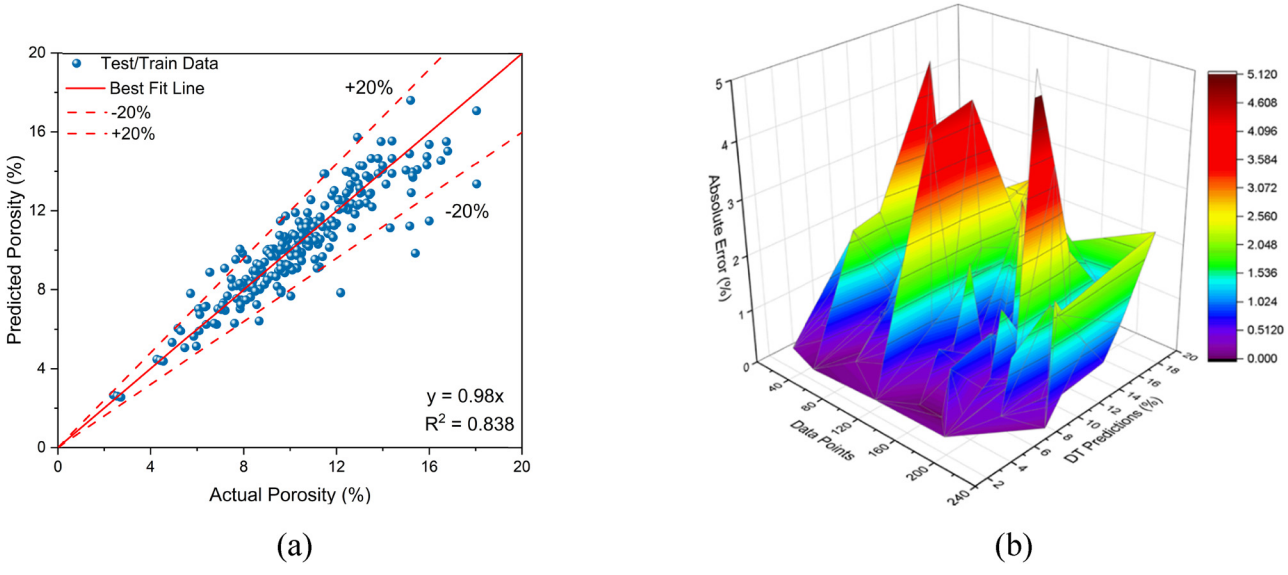

Figure 10: DT models outcomes: (a) R^2 graph and (b) error tracing.

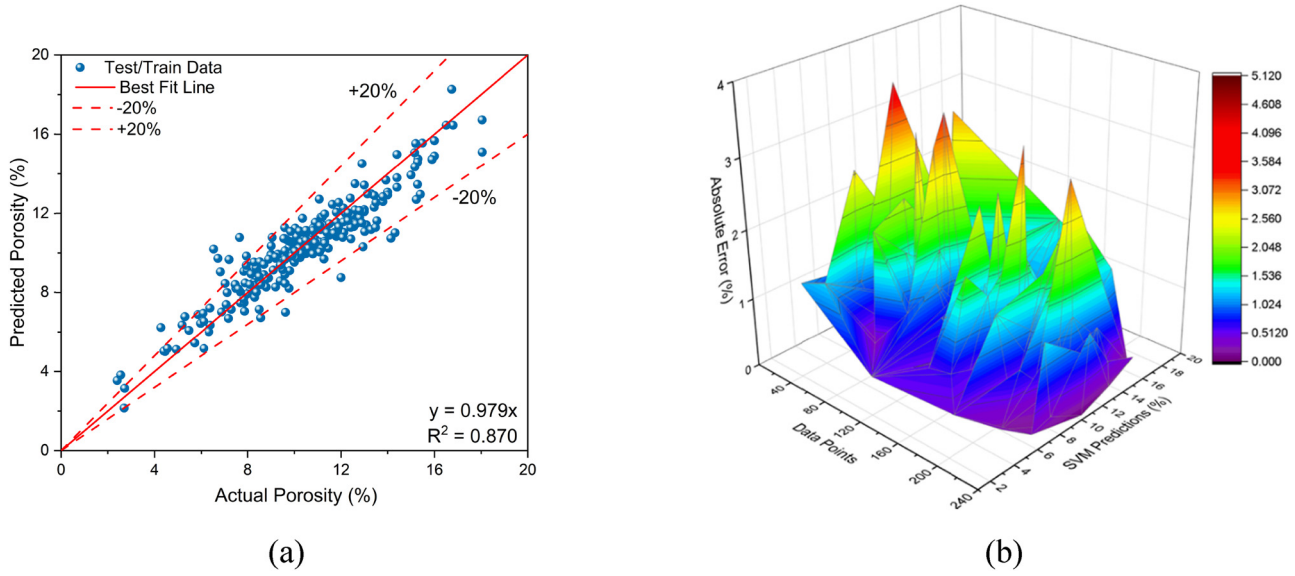

Figure 11: SVM model outcomes: (a) R^2 graph and (b) error tracing.

limit, and mean limit error readings of 3.64, 0.01, and 0.80%, respectively. Overall, 71.25% of the errors observed were found to be less than 1%, while 26.67% were within the range of 1–3%. The remaining 2.08% of the error values exceeded 3%. The superior accuracy of the SVM algorithm, in contrast to the DT algorithm, is additionally substantiated by the presence of lower error values.

4.4 Adaptive boosting outcomes

Figure 12(a) depicts the correlation between the experimental findings of the AdaBoost and the projected outcomes. The R^2 value of 0.914 for this model suggests that it possesses a higher degree of accuracy in predicting responses. AdaBoost shows higher accuracy than SVM, DT,

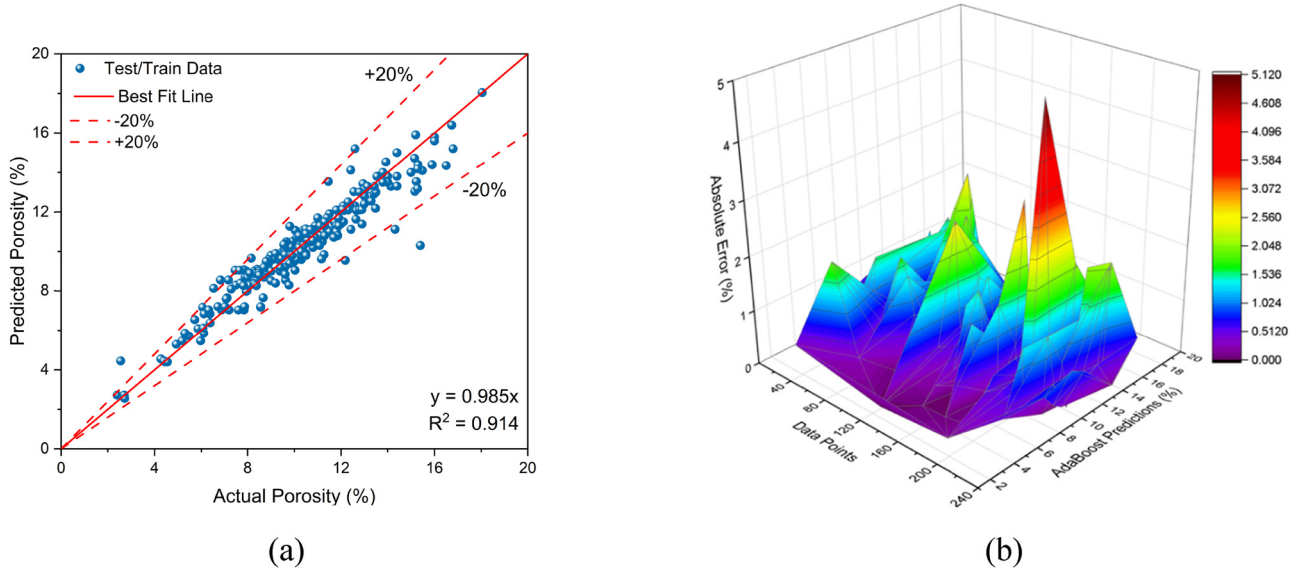

Figure 12: AdaBoost model outcomes: (a) R^2 graph and (b) error tracing.

Table 7: Error evaluation of models

Models	MAE (%)	RMSE (%)	R^2	MAPE (%)	OBJ	SI
AdaBoost	0.617	0.858	0.914	6.3	0.771	0.083
SVM	0.796	1.059	0.870	8.5	0.992	0.103
DT	0.824	1.166	0.838	7.9	1.083	0.114
LR	1.128	1.410	0.763	12.6	1.440	0.136

and LR due to its iterative adaptation and aggregation of several weak classifiers. This collective approach effectively improves model performance by mitigating misclassification errors. Figure 12(b) displays a graphical depiction of the distribution of observed, forecasted, and error values within the AdaBoost model. Within the training set, the highest upper boundary, lowest upper boundary, and average boundary values were recorded as 5.10, 0.00, and 0.62%, respectively. Nevertheless, 83.33% of the inaccurate approximations were below 1%, whereas only 0.83% were above 3%. The AdaBoost model demonstrated greater precision in forecasting concrete's porosity relative to the LR, DT, and SVM models, as supported by the comparison of R^2 values and error distributions. All of the model's R^2 values and error percentages fell within acceptable levels, indicating improved predictive outcomes.

4.5 Statistical error analysis

The outcome of this study contributed to the creation of ML algorithms, which were optimized to predict the porosity of concrete and subsequently compare the performance of these models. Table 7 presents the statistical deviations observed between the anticipated and experimental results. The statistical error analysis reveals a strong association between the observed and projected responses for the AdaBoost model, demonstrating a high accuracy level in forecasting the porosity of concrete. The AdaBoost model has exceptional performance, as demonstrated by its R^2 score of 0.914 and the following evaluation metrics: RMSE (0.858%), SI (0.083), MAE (0.617%), OBJ (0.771%), and MAPE (6.3%). Nevertheless, the LR model demonstrates the lowest level of accuracy, as specified by an R^2 score of 0.763 and errors in terms of RMSE, SI, MAE, OBJ, and MAPE, which amount to 1.410, 0.136, 1.128, 1.440, and 12.6%, respectively. The violin plot shown in Figure 13 demonstrates the error propagation of ML models. The MAE of the AdaBoost model is 0.617%, lower than 0.796% for SVM, 0.824% for DT, and 1.128% for the LR model. Thus, the violin plot testifies to the boosting performance of the AdaBoost model against DT, SVM, and LR models in terms of forecasting the porosity of concrete.

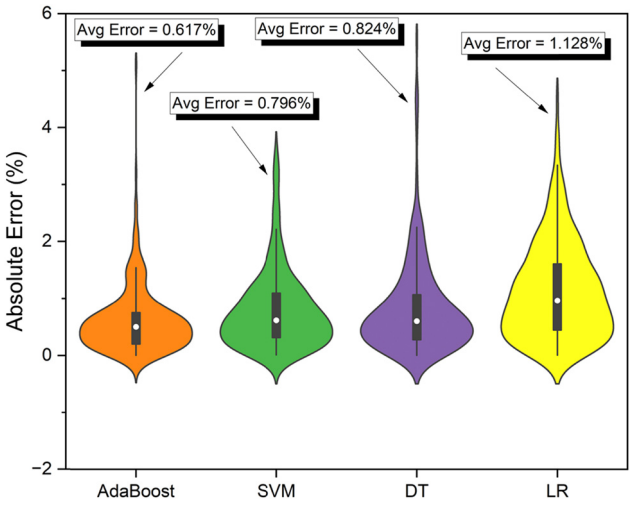

Figure 13: Violin graph for various developed models.

4.6 Feature importance evaluation

To forecast the porosity of concrete, this research employed CA/FA, SP, w/b, GGBFS, CT, fly ash, and binder as input features. Further, this study employed the RReliefF algorithm to order the input features according to their importance while estimating the output target. Figure 14 presents the findings of the RReliefF test. According to this radial plot, the binder content, w/b, and GGBFS percentage significantly influence the porosity of concrete with the RReliefF factors of 0.167, 0.151, and 0.150, respectively. However, SPs and curing time have the most negligible effect on the output target (porosity), as suggested by the RReliefF factors of 0.085 and 0.084, respectively. The porosity of cementitious composite is greatly affected by the amount of binder present, as the binder is responsible for the hydration process that bonds the aggregate fragments together. The quantity and composition of the binder are key factors in determining the size and arrangement of the hydration products, such as CSH, that occupy the voids in the concrete structure. The calcium-to-silica ratio (C/S ratio) in CSH has an impact on the pore structure and permeability [78]. Inadequate binder causes incomplete hydration, leading to an increase in capillary pores and increased porosity. On the other hand, an ideal quantity of binder decreases porosity by guaranteeing a more compact microstructure through efficient filling of pores [79]. Prior research has also demonstrated that fluctuations in the amount of binder significantly impact concrete's mechanical properties and permeability, highlighting the significance of regulating the ratio of binder to water to attain the necessary porosity and strength [80]. Since the binder and w/b ratio are the controlling factors affecting the concrete porosity, while preparing a concrete mix, one can prioritize these variables

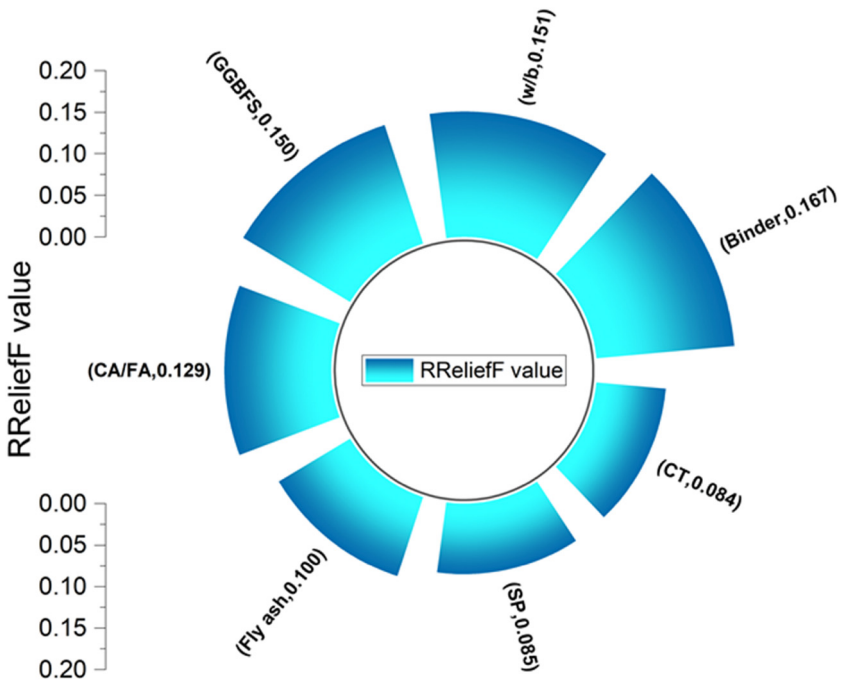

Figure 14: Radial plot for feature importance.

and compute the other quantities. Moreover, optimal concrete mix designs can be achieved by modifying the quantity of binder to attain the required porosity level.

4.7 Sieve diagram

Figure 15(a)–(d) demonstrates the sieve diagram for DT, LR, SVM, and AdaBoost models. The sieve diagram is a graphic technique applied to visually characterize frequencies in a two-way contingencies table and assess them concerning the expected frequencies based on a presumption of independence. The presented visualization depicts rectangles whose areas are proportionate to the expected frequency, while the quantity of squares represents the experimental frequency within each rectangle. The shading density visually represents the distinction between actual and forecasted frequency, which is proportional to the standard Pearson residual (χ^2). This shading utilizes color to show whether the divergence from independence is red (−ive) or blue (+ive). The χ^2 value for AdaBoost is 375.09, which is higher than 208.80 for DT, 319.29 for SVM, and 99.47 for the LR model. The higher χ^2 suggests a strong correlation between actual and anticipated outcomes. Hence, AdaBoost shows an excellent correlation of experimental porosity with predicted porosity.

5 Discussion

The current research employed a random split of the porosity database, with a ratio of 70% for training and 30% for testing. This partition was used for all ML models, including DT, SVM, AdaBoost, and LR. The intent was to evaluate and contrast the predictive accuracy of these techniques and study the applicability of these optimized models for calculating the porosity of concrete without any lab experimentation. The correlation coefficient of all developed models showed acceptable performance, while AdaBoost outperformed SVM, LR, and DT models in terms of accuracy and predictive capabilities. AdaBoost shows higher accuracy than SVM, DT, and LR due to its iterative adaptation and aggregation of several weak classifiers. This collective approach effectively improves model performance by mitigating misclassification errors. AdaBoost is an ML algorithm that focuses on complex data points and iteratively enhances its predictive performance by allocating more weights to previously misclassified instances. The effectiveness of this technique lies in its adaptability and ensemble nature, which allows it to effectively capture complicated decision boundaries and address difficulties related to overfitting. Additionally, it demonstrates versatility by accommodating different types of base classifiers. In addition, AdaBoost exhibits less sensitivity to feature scaling, has a simplicity of implementation, and has the capacity to

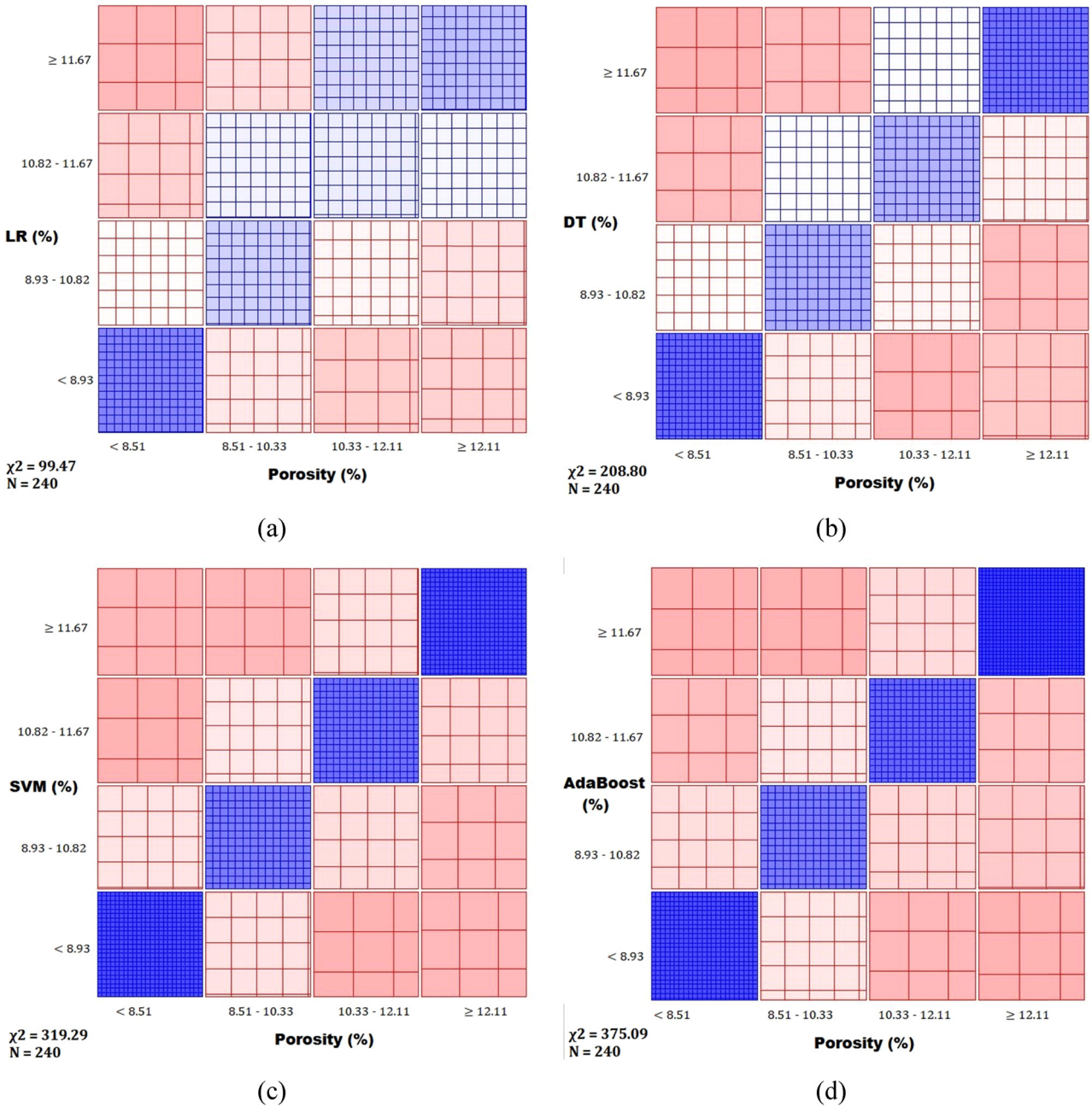

Figure 15: Sieve diagrams: (a) LR model, (b) DT model, (c) SVM model, and (d) AdaBoost model.

effectively handle non-linear data, which collectively contributes to its strong performance in many datasets and applications. Compared to the LR model, which is primarily based on linear assumption, the DT and SVM models perform well as both models can trace out intrinsic correlation among the variables. However, the efficiency of these algorithms is highly dependent on the database. Any potential outlier in the database yields false results in computation modeling.

The approach described above exhibits promising potential for several applications, such as optimizing mix compositions to enhance the performance-driven design of concrete structures and mitigating the adverse ecological effects associated with concrete production. Furthermore, the ML technique under consideration can be utilized to provide empirical predictions regarding the specific durability characteristics of concrete, including but not limited to the water and gas permeability, chloride diffusion coefficient,

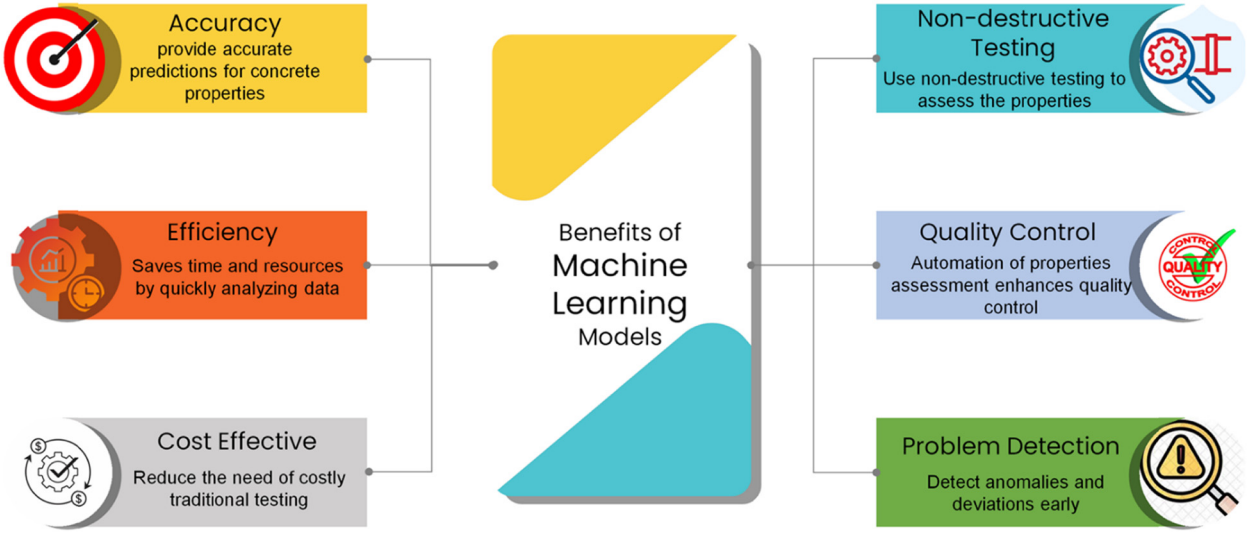

Figure 16: Applications of ML for academics and industry.

electrical resistivity, and degree of cement hydration. Subsequent research endeavors should focus on developing a dependable and balanced database encompassing concrete porosity, which can be utilized to train the ML model. The integration of diverse ML techniques can achieve the enhancement of forecasting accuracy. The applications of ML for industry and academics are described in Figure 16. However, some constraints are associated with utilizing these ML techniques. The exploitation of large, superior databases is crucial for ensuring the effectiveness of training ML models. However, the process of obtaining a comprehensive and diversified database that includes specific compositions of cementitious compounds and their accompanying porosity values may present significant difficulties. The lack of available data has the potential to compromise the accuracy and relevance of the algorithm. The algorithm's ability to make accurate predictions for unfamiliar mixtures or unusual processing parameters may be compromised if the training data lacks coverage of the full range of compositions of cementitious compounds or unique situations of interest.

6 Conclusions

This research introduced an innovative machine learning (ML) framework that demonstrated exceptional precision and adaptation performance in predicting the porosity of concrete. ML techniques include DT, AdaBoost, LR, and SVM. The models were subjected to intensive training and testing utilizing data sourced from reputable scholarly publications. Statistical analyses have demonstrated that ML algorithms exhibit superior performance compared to conventional approaches, characterized by enhanced generalization capabilities and greater reliability. The study main findings are as follows:

1) The ML procedures suggested in this study exhibited a notable level of efficiency and generalization capability when employed for predicting the porosity of concrete. The correlation coefficient (R^2) for AdaBoost was observed to be 0.914. Meanwhile, R^2 values for DT, SVM, and LR were 0.838, 0.870, and 0.763, respectively.
2) The results of RMSE (0.858), MAPE (6.3), SI (0.083), MAE (0.617), and OBJ (0.771) for the AdaBoost-model confirmed the superior efficiency than the LR-model (1.410, 12.6, 0.136, 1.128, and 1.440), the DT-model (1.166, 7.9, 0.114, 0.824, and 1.083), and the SVM model (1.059, 8.5, 0.013, 0.796, and 0.992).
3) As evident from the R^2 score and minimum observed errors, the AdaBoost algorithm exhibited reliable accuracy and predictive capabilities against SVM, LR, and DT for computing the porosity of concrete.
4) The outcomes of RReliefF analysis demonstrated the comparative importance of each independent variable, with the following order of significance: binder > water/binder > slag > coarse/fine particles > fly ash > superplasticizer > curing time.
5) The AdaBoost model's 375.09 chi-square value (χ^2) showed a significant correlation of inputs with the target variable. However, the χ^2 values for DT, SVM, and LR were observed to be 208.50, 319.29, and 99.47, respectively.

The research findings suggest that the formulated models yield higher accuracy and precision. However,

the use of a diverse dataset with a greater number of input features can lead to the formation of a universal ML model capable of capturing the intrinsic correlation among the variables. Further studies can be conducted to generate a dataset that includes the chemical composition of SCMs and curing conditions to create a universal ML model for estimating the porosity of concrete.

Acknowledgments: The authors acknowledge the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU241889). The authors extend their appreciation for the financial support that made this study possible.

Funding information: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU241889].

Author contributions: H.H.: conceptualization, methodology, visualization, formal analysis, and writing – original draft. M.N.A.: funding acquisition, supervision, project administration, investigation, writing – reviewing, and editing. S.A.K.: data acquisition, software, methodology, validation, and writing – original draft. K.K.: conceptualization, resources, formal analysis, writing – reviewing, and editing. M.T.Q.: visualization, project administration, investigation, writing – reviewing, and editing. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Claisse, P. A., J. G. Cabrera, and D. N. Hunt. Measurement of porosity as a predictor of the durability performance of concrete with and without condensed silica fume. *Advances in Cement Research*, Vol. 13, 2001, pp. 165–174.

[2] Linares-Alemparte, P., C. Andrade, and D. Baza. Porosity and electrical resistivity-based empirical calculation of the oxygen diffusion coefficient in concrete. *Construction and Building Materials*, Vol. 198, 2019, pp. 710–717.

[3] Song, H.-W. and S.-J. Kwon. Permeability characteristics of carbonated concrete considering capillary pore structure. *Cement and Concrete Research*, Vol. 37, 2007, pp. 909–915.

[4] Simčič, T., S. Pejovnik, G. De Schutter, and V. B. Bosiljkov. Chloride ion penetration into fly ash modified concrete during wetting–drying cycles. *Construction and Building Materials*, Vol. 93, 2015, pp. 1216–1223.

[5] Bertolini, L., B. Elsener, P. Pedeferri, E. Redaelli, and R. B. Polder. *Corrosion of steel in concrete: prevention, diagnosis, repair*, John Wiley & Sons, Weinheim, Germany, 2013.

[6] Cao, C. Prediction of concrete porosity using machine learning. *Results in Engineering*, Vol. 17, 2023, id. 100794.

[7] Därr, G. M. and U. Ludwig. Determination of permeable porosity. *Matériaux et Construction*, Vol. 6, 1973, pp. 185–190.

[8] Winslow, D. and D. Liu. The pore structure of paste in concrete. *Cement and Concrete Research*, Vol. 20, 1990, pp. 227–235.

[9] Hansen, T. C. Physical structure of hardened cement paste. A classical approach. *Materials and Structures*, Vol. 19, 1986, pp. 423–436.

[10] Basheer, L., P. A. M. Basheer, and A. E. Long. Influence of coarse aggregate on the permeation, durability and the microstructure characteristics of ordinary Portland cement concrete. *Construction and Building Materials*, Vol. 19, 2005, pp. 682–690.

[11] Ahmad, S., A. K. Azad, and K. F. Loughlin. Effect of the key mixture parameters on tortuosity and permeability of concrete. *Journal of Advanced Concrete Technology*, Vol. 10, 2012, pp. 86–94.

[12] Thomas, M. D. A. and P. B. Bamforth. Modelling chloride diffusion in concrete: Effect of fly ash and slag. *Cement and Concrete Research*, Vol. 29, 1999, pp. 487–495.

[13] Papadakis, V. G. Effect of supplementary cementing materials on concrete resistance against carbonation and chloride ingress. *Cement and Concrete Research*, Vol. 30, 2000, pp. 291–299.

[14] Kakasor Ismael Jaf, D., P. Ismael Abdulrahman, A. Salih Mohammed, R. Kurda, S. M. A. Qaidi, and P. G. Asteris. Machine learning techniques and multi-scale models to evaluate the impact of silicon dioxide (SiO2) and calcium oxide (CaO) in fly ash on the compressive strength of green concrete. *Construction and Building Materials*, Vol. 400, 2023, id. 132604.

[15] Miller, S. A., P. J. M. Monteiro, C. P. Ostertag, and A. Horvath. Concrete mixture proportioning for desired strength and reduced global warming potential. *Construction and Building Materials*, Vol. 128, 2016, pp. 410–421.

[16] Song, H.-W. and V. Saraswathy. Studies on the corrosion resistance of reinforced steel in concrete with ground granulated blast-furnace slag – An overview. *Journal of Hazardous Materials*, Vol. 138, 2006, pp. 226–233.

[17] Thomas, M. D. A. and J. D. Matthews. The permeability of fly ash concrete. *Materials and Structures*, Vol. 25, 1992, pp. 388–396.

[18] Hassan, K. E., J. G. Cabrera, and R. S. Maliehe. The effect of mineral admixtures on the properties of high-performance concrete. *Cement and Concrete Composites*, Vol. 22, 2000, pp. 267–271.

[19] Papadakis, V. G. Effect of fly ash on Portland cement systems: Part I. Low-calcium fly ash. *Cement and Concrete Research*, Vol. 29, 1999, pp. 1727–1736.

[20] Papadakis, V. G. Effect of fly ash on Portland cement systems: Part II. High-calcium fly ash. *Cement and Concrete Research*, Vol. 30, 2000, pp. 1647–1654.

[21] Salih, A., S. Rafiq, P. Sihag, K. Ghafor, W. Mahmood, and W. Sarwar. Systematic multiscale models to predict the effect of high-volume fly ash on the maximum compression stress of cement-based mortar at various water/cement ratios and curing times. *Measurement*, Vol. 171, 2021, id. 108819.

[22] Mohammed, A., R. Kurda, D. J. Armaghani, and M. Hasanipanah. Prediction of compressive strength of concrete modified with fly

ash: Applications of neuro-swarm and neuro-imperialism models. *Computers and Concrete*, Vol. 27, 2021, pp. 489–512.
[23] Piro, N. S., A. S. Mohammed, and S. M. Hamad. The impact of GGBS and ferrous on the flow of electrical current and compressive strength of concrete. *Construction and Building Materials*, Vol. 349, 2022, id. 128639.
[24] Piro, N. S., A. S. Mohammed, and S. M. Hamad. Evaluate and predict the resist electric current and compressive strength of concrete modified with GGBS and steelmaking slag using mathematical models. *Journal of Sustainable Metallurgy*, Vol. 9, 2023, pp. 194–215.
[25] Yaman, I. O., H. M. Aktan, and N. Hearn. Active and non-active porosity in concrete part II: evaluation of existing models. *Materials and Structures*, Vol. 35, 2002, pp. 110–116.
[26] Kim, Y.-Y., K.-M. Lee, J.-W. Bang, and S.-J. Kwon. Effect of W/C ratio on durability and porosity in cement mortar with constant cement amount. *Advances in Materials Science and Engineering*, Vol. 2014, 2014, id. 273460.
[27] Zanovello, M., R. Baldusco, V. M. John, and S. C. Angulo. Strength-porosity correlation and environmental analysis of recycled Portland cement. *Resources, Conservation and Recycling*, Vol. 190, 2023, id. 106763.
[28] Schindler, A. K. and K. J. Folliard. Heat of hydration models for cementitious materials. *ACI Materials Journal*, Vol. 102, 2005, id. 24.
[29] Riding, K. A., J. L. Poole, K. J. Folliard, M. C. G. Juenger, and A. K. Schindler. Modeling hydration of cementitious systems. *ACI Materials Journal*, Vol. 109, 2012, pp. 225–234.
[30] Shafikhani, M. and S. E. Chidiac. A holistic model for cement paste and concrete chloride diffusion coefficient. *Cement and Concrete Research*, Vol. 133, 2020, id. 106049.
[31] Khan, M. I. Permeation of high performance concrete. *Journal of Materials in Civil Engineering*, Vol. 15, 2003, pp. 84–92.
[32] Faraz, M. I., S. U. Arifeen, M. N. Amin, A. Nafees, F. Althoey, and A. Niaz. A comprehensive GEP and MEP analysis of a cement-based concrete containing metakaolin. In *Structures*, Elsevier, 2023, pp. 937–948.
[33] Arifeen, S. U., M. N. Amin, W. Ahmad, F. Althoey, M. Ali, B. S. Alotaibi, et al. A comparative study of prediction models for alkali-activated materials to promote quick and economical adaptability in the building sector. *Construction and Building Materials*, Vol. 407, 2023, id. 133485.
[34] Mohammed, A., L. Burhan, K. Ghafor, W. Sarwar, and W. Mahmood. Artificial neural network (ANN), M5P-tree, and regression analyses to predict the early age compression strength of concrete modified with DBC-21 and VK-98 polymers. *Neural Computing and Applications*, Vol. 33, 2021, pp. 7851–7873.
[35] Ahmed, H. U., A. S. Mohammed, and A. A. Mohammed. Proposing several model techniques including ANN and M5P-tree to predict the compressive strength of geopolymer concretes incorporated with nano-silica. *Environmental Science and Pollution Research*, Vol. 29, 2022, pp. 71232–71256.
[36] Nazar, S., J. Yang, M. N. Amin, K. Khan, M. F. Javed, and F. Althoey. Formulation of estimation models for the compressive strength of concrete mixed with nanosilica and carbon nanotubes. *Developments in the Built Environment*, Vol. 13, 2023, id. 100113.
[37] Ling, H., C. Qian, W. Kang, C. Liang, and H. Chen. Combination of support vector machine and K-Fold cross validation to predict compressive strength of concrete in marine environment. *Construction and Building Materials*, Vol. 206, 2019, pp. 355–363.
[38] Pereira, L., L. Godinho, F. G. Branco, and P. da Venda Oliveira. A machine-learning based approach to estimate acoustic macroscopic parameters of porous concrete. *Construction and Building Materials*, Vol. 426, 2024, id. 136075.
[39] Sathiparan, N., S. H. Wijekoon, P. Jeyananthan, and D. N. Subramaniam. Soft computing to predict the porosity and permeability of pervious concrete based on mix design and ultrasonic pulse velocity. *International Journal of Pavement Engineering*, Vol. 25, 2024, id. 2337916.
[40] Wu, Y., R. Pieralisi, F. G. Sandoval, R. D. López-Carreño, and P. Pujadas. Optimizing pervious concrete with machine learning: Predicting permeability and compressive strength using artificial neural networks. *Construction and Building Materials*, Vol. 443, 2024, id. 137619.
[41] Boukhatem, B., R. Rebouh, A. Zidol, M. Chekired, and A. Tagnit-Hamou. An intelligent hybrid system for predicting the tortuosity of the pore system of fly ash concrete. *Construction and Building Materials*, Vol. 205, 2019, pp. 274–284.
[42] Le, B.-A., V.-H. Vu, S.-Y. Seo, B.-V. Tran, T. Nguyen-Sy, M.-C. Le, et al. Predicting the compressive strength and the effective porosity of pervious concrete using machine learning methods. *KSCE Journal of Civil Engineering*, Vol. 26, 2022, pp. 4664–4679.
[43] Mahjoubi, S., W. Meng, and Y. Bao. Auto-tune learning framework for prediction of flowability, mechanical properties, and porosity of ultra-high-performance concrete (UHPC). *Applied Soft Computing*, Vol. 115, 2022, id. 108182.
[44] Cao, C. Machine learning-based prediction of porosity for concrete containing supplementary cementitious materials. *arXiv preprint arXiv:211207353*, Vol. 17, 2021, id. 100794.
[45] Cheng, A.-S., T. Yen, Y.-W. Liu, and Y.-N. Sheen. Relation between porosity and compressive strength of slag concrete, In: *Structures Congress 2008: Crossing Borders*, 2008, pp. 1–8.
[46] Al-Amoudi, O. S. B., I. M. Asi, and M. Maslehuddin. Performance and correlation of the properties of fly ash cement concrete. *Cement, Concrete, and Aggregates*, Vol. 18, 1996, pp. 71–77.
[47] Shafiq, N., M. F. Nuruddin, and I. Kamaruddin. Comparison of engineering and durability properties of fly ash blended cement concrete made in UK and Malaysia. *Advances in Applied Ceramics*, Vol. 106, 2007, pp. 314–318.
[48] Van den Heede, P., E. Gruyaert, and N. De Belie. Transport properties of high-volume fly ash concrete: Capillary water sorption, water sorption under vacuum and gas permeability. *Cement and Concrete Composites*, Vol. 32, 2010, pp. 749–756.
[49] Younsi, A., P. Turcry, E. Rozière, A. Aït-Mokhtar, and A. Loukili. Performance-based design and carbonation of concrete with high fly ash content. *Cement and Concrete Composites*, Vol. 33, 2011, pp. 993–1000.
[50] Ahmad, S. and A. K. Azad. An exploratory study on correlating the permeability of concrete with its porosity and tortuosity. *Advances in Cement Research*, Vol. 25, 2013, pp. 288–294.
[51] Li, Y., G. Wang, M. N. Amin, A. Khan, M. T. Qadir, and S. U. Arifeen. Towards improved flexural behavior of plastic-based mortars: An experimental and modeling study on waste material incorporation. *Materials Today Communications*, Vol. 40, 2024, id. 109391.
[52] Zhou, J., Q. Tian, S. Nazar, and J. Huang. Hyper-tuning gene expression programming to develop interpretable prediction models for the strength of corncob ash-modified geopolymer concrete. *Materials Today Communications*, Vol. 38, 2024, id. 107885.
[53] Kewalramani, M. and A. Khartabil. Porosity evaluation of concrete containing supplementary cementitious materials for durability assessment through volume of permeable voids and water immersion conditions. *Buildings*, Vol. 11, 2021, id. 378.

[54] Tian, Q., Y. Lu, J. Zhou, S. Song, L. Yang, T. Cheng, et al. Supplementary cementitious materials-based concrete porosity estimation using modeling approaches: A comparative study of GEP and MEP. *Reviews on Advanced Materials Science*, Vol. 63, 2024, id. 20230189.
[55] Sun, H., M. N. Amin, M. T. Qadir, S. U. Arifeen, B. Iftikhar, and F. Althoey. Investigating the effectiveness of carbon nanotubes for the compressive strength of concrete using AI-aided tools. *Case Studies in Construction Materials*, Vol. 20, 2024, id. e03083.
[56] Kotsiantis, S. B. Decision trees: a recent overview. *Artificial Intelligence Review*, Vol. 39, 2013, pp. 261–283.
[57] Erdal, H. I. Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. *Engineering Applications of Artificial Intelligence*, Vol. 26, 2013, pp. 1689–1697.
[58] Karbassi, A., B. Mohebi, S. Rezaee, and P. Lestuzzi. Damage prediction for regular reinforced concrete buildings using the decision tree algorithm. *Computers & Structures*, Vol. 130, 2014, pp. 46–56.
[59] El Asri, Y., M. B. Aicha, M. Zaher, and A. H. Alaoui. Prediction of compressive strength of self-compacting concrete using four machine learning technics. *Materials Today: Proceedings*, Vol. 57, 2022, pp. 859–866.
[60] Nazar, S., J. Yang, W. Ahmad, M. F. Javed, H. Alabduljabbar, and A. F. Deifalla. Development of the new prediction models for the compressive strength of nanomodified concrete using novel machine learning techniques. *Buildings*, Vol. 12, 2022, id. 2160.
[61] Liang, H. and W. Song. Improved estimation in multiple linear regression models with measurement error and general constraint. *Journal of Multivariate Analysis*, Vol. 100, 2009, pp. 726–741.
[62] Chou, J.-S., C.-F. Tsai, A.-D. Pham, and Y.-H. Lu. Machine learning in concrete strength simulations: Multi-nation data analytics. *Construction and Building Materials*, Vol. 73, 2014, pp. 771–780.
[63] Vapnik, V. *The nature of statistical learning theory*, Springer Science & Business Media, New York, 1999.
[64] Smola, A. J. and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, Vol. 14, 2004, pp. 199–222.
[65] Freund, Y., R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, Vol. 14, 1999, id. 1612.
[66] Wang, C., S. Xu, and J. Yang. Adaboost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement. *Sensors*, Vol. 21, 2021, id. 5682.
[67] Ying, C., M. Qi-Guang, L. Jia-Chen, and G. Lin. Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, Vol. 39, 2013, pp. 745–758.
[68] Schapire, R. E. Explaining adaboost. In *Empirical inference: Festschrift in honor of vladimir N. Vapnik*, Springer, Berlin Heidelberg, 2013, pp. 37–52.
[69] Nasir Amin, M., B. Iftikhar, K. Khan, M. Faisal Javed, A. Mohammad AbuArab, and M. Faisal Rehman. Prediction model for rice husk ash concrete using AI approach: Boosting and bagging algorithms. *Structures*, Vol. 50, 2023, pp. 745–757.
[70] Rebba, R. and S. Mahadevan. Validation of models with multivariate output. *Reliability Engineering & System Safety*, Vol. 91, 2006, pp. 861–871.
[71] Zhang, H., X. Li, M. N. Amin, A. A. Alawi Al-Naghi, S. U. Arifeen, F. Althoey, et al. Analyzing chloride diffusion for durability predictions of concrete using contemporary machine learning strategies. *Materials Today Communications*, Vol. 38, 2024, id. 108543.
[72] Ahmad, A., W. Ahmad, F. Aslam, and P. Joyklad. Compressive strength prediction of fly ash-based geopolymer concrete *via* advanced machine learning techniques. *Case Studies in Construction Materials*, Vol. 16, 2022, id. e00840.
[73] Cao, Q., X. Yuan, M. N. Amin, W. Ahmad, F. Althoey, and F. Alsharari. A soft-computing-based modeling approach for predicting acid resistance of waste-derived cementitious composites. *Construction and Building Materials*, Vol. 407, 2023, id. 133540.
[74] Kira, K. and L. A. Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, Elsevier, 1992, pp. 249–256.
[75] Sumant, A. S. and D. Patil. Ensemble feature subset selection: Integration of symmetric uncertainty and chi-square techniques with RReliefF. *Journal of The Institution of Engineers (India): Series B*, Vol. 103, 2022, pp. 831–844.
[76] Kononenko, I. *Estimating attributes: Analysis and extensions of RELIEF*, Springer European conference on machine learning, Berlin, Heidelberg, 1994, pp. 171–182.
[77] Robnik-Šikonja, M. and I. Kononenko. *An adaptation of Relief for attribute estimation in regression*, Machine learning: Proceedings of the fourteenth international conference (ICML'97), 1997, pp. 296–304.
[78] Li, P., D. Su, S. Wang, and Z. Fan. Influence of binder composition and concrete pore structure on chloride diffusion coefficient in concrete. *Journal of Wuhan University of Technology-Mater Sci Ed*, Vol. 26, 2011, pp. 160–164.
[79] Lindh, P. and P. Lemenkova. Effects of Water – Binder ratio on strength and seismic behavior of stabilized soil from Kongshavn, Port of Oslo. *Sustainability*, Vol. 15, 2023.
[80] Baldovino, J. D., Y. E. Nuñez de la Rosa, and O. P. Calabokis. Effect of porosity/binder index on strength, stiffness and microstructure of cemented clay: The impact of sustainable development geomaterials. *Materials*, Vol. 17, 2024, id. 921.