symmetry

MDPI

Article

Evaluation of Classification for Project Features with Machine Learning Algorithms

Ching-Lung Fan

Department of Civil Engineering, The Republic of China Military Academy, No. 1, Weiwu Rd., Fengshan, Kaohsiung 830, Taiwan; p93228001@ntu.edu.tw; Tel.: +886-7-7456290

Abstract: Due to the asymmetry of project features, it is difficult for project managers to make a reliable prediction of the decision-making process. Big data research can establish more predictions through the results of accurate classification. Machine learning (ML) has been widely applied for big data analytic and processing, which includes model symmetry/asymmetry of various prediction problems. The purpose of this study is to achieve symmetry in the developed decision-making solution based on the optimal classification results. Defects are important metrics of construction management performance. Accordingly, the use of suitable algorithms to comprehend the characteristics of these defects and train and test massive data on defects can conduct the effectual classification of project features. This research used 499 defective classes and related features from the Public Works Bid Management System (PWBMS). In this article, ML algorithms, such as support vector machine (SVM), artificial neural network (ANN), decision tree (DT), and Bayesian network (BN), were employed to predict the relationship between three target variables (engineering level, project cost, and construction progress) and defects. To formulate and subsequently cross-validate an optimal classification model, 1015 projects were considered in this work. Assessment indicators showed that the accuracy of ANN for classifying the engineering level is 93.20%, and the accuracy values of SVM for classifying the project cost and construction progress are 85.32% and 79.01%, respectively. In general, the SVM yielded better classification results from these project features. This research was based on an ML algorithm evaluation system for buildings as a classification model for project features with the goal of aiding project managers to comprehend defects.

Keywords: support vector machines; artificial neural network; decision trees; Bayesian network; machine learning; defects

Citation: Fan, C.-L. Evaluation of Classification for Project Features with Machine Learning Algorithms. *Symmetry* **2022**, *14*, 372. https://doi.org/10.3390/sym14020372

Academic Editor: Cengız Kahraman

Received: 22 November 2021
Accepted: 10 February 2022
Published: 13 February 2022

1. Introduction

Defects are the focus of quality management and indicators of project performance. Meng [1] studied the performance information of 103 projects and found that there were 90 quality defects, 37 delayed times, and 26 overspending costs. The quality defect items contributed to 87.4% of all the items. Thus, defects are one of the main factors for poor construction performance. Early scholars have studied defects with a focus on the classification of defects and the types of defects [2–6]. However, the analytical methods used lack the capability of automatic data exploration, and the evaluation process is complex, time-consuming, and often requires professionals to set the appropriate parameters to obtain the correct results. The machine learning (ML) algorithm can compensate for the shortcomings of these methods. By using nontraditional analysis methods and using a large database for importing the ML algorithm, it is possible to classify project features and defects in a simpler model. Due to the frequent occurrence of many defects during a construction process, ML can be suitably used in the construction industry to enable a project manager to clearly determine the relationship between defects and project features.

Currently, several scholars have used artificial intelligence (AI) and ML, combined with big data for defect assessment, or have used a database as a method to estimate

defects [7–10]. These studies combined ML algorithms and established decision support for expert systems provide an efficient means to solve problems, quickly determine the ability of construction defects, and achieve a certain level of predictability. The performance of a model established using machine learning relies to a great degree on the size and density of the obtainable training dataset [11]. Often, inadequate data are available owing to the randomness of the data distribution and incompatible inspection record, which happens because of inconsistent inspection processes [12]. Therefore, the challenge of a defect analysis model is that strict data collection agreements are required before application [13]. Macarulla et al. [14] represented that defect data are structurally unreasonable or is not readily obtainable; thence, data analysis is difficult. Das and Chew [15] specified that a scientific rating system should be built using a defect database and conducting a systematic grading to explicitly define the influence of the defects.

Taiwan's government units systematically record construction inspection data in the Public Works Bid Management System (PWBMS) through the inspection mechanism. After, data analyses were executed, and defect improvement approaches were conducted to upgrade the quality of public works and construction performance. To date, a total of 499 defects identified by inspection were classified into four categories: construction management, work quality, program, and design. The feature data of defects used in the current study were sourced from the PWBMS. The defects were gained by committee (scholars and experts) in construction site inspections adopted official standardized forms. Therefore, the inspection criteria were consistent and unprejudiced. In addition, the database contained 27 years of defect data, and it was extremely large.

The use of considerable amounts of data to train machine models enables the avoidance of statistical methods and the limitations of frameworks. By replacing sample analysis with big data analysis, observing the relationship among data points, recognizing patterns that are previously difficult to realize, and applying the value of new thinking become possible. Consequently, data cease to be treated as afterthoughts that are subsequently arranged and applied; instead, they are regarded as tools for building and exploring problem domains. Accordingly, they enable the creation and development of technical abilities and expertise in the field of domain knowledge management [16]. The purpose of big data analysis is to discover knowledge, predict results, and support decision-making to create a competitive advantage [17].

Since the introduction of the construction inspection system in Taiwan, considerable construction features of big data related to public projects have accumulated. Currently, academic scholars and the industry have been focusing on how to mine big data and analyze the results correctly. Defects indicate a project's construction quality, and the relationship between these defects and the features related to the project is an issue that requires further analysis. Thus, by obtaining useful rules and knowledge that can be derived from an inspection database, the relationship between the defects and related features can be comprehended, thereby reducing or removing the risk of future defects.

Because ML can process multidimensional data or information and explore the association between multiple variables while conducting exploration of data involving statistics and large data sets, it is suitable for big data analysis. ML has been widely applied for big data analytic and processing, which includes model symmetry/asymmetry of various pre-diction problems. Applying ML algorithms to analyze big data can reveal modes that were not obvious. Sometimes, unsuspected correlations or new trends may be discovered, thereby leading to a better solving of the problem.

Based on the above study background and the characteristics of ML, this study combined massive data of defects and ML to obtain hidden information that has not been explored or considered valuable in the past or to discover patterns and rules. Due to the asymmetry of project features, it is difficult for project managers to make a reliable prediction of the decision-making process. Big data research can establish more predictions through the results of accurate classification. The purpose of this research is to employ ML to identify the hidden connection of project features from a large amount of inspection data

and achieve symmetry in the developed decision-making solution based on the optimal classification results. The relationship between target variables (project features) and defects is predicted. A classification model for optimizing project features has been established to provide construction managers with assistant means for comprehending defects. Managers can be enabled to achieve correct decisions to improve project management strategy and construction performance.

2. Construction Inspection Data and Machine Learning

2.1. Features of Construction Data

This research utilized the construction inspection data from PWBMS (1993 to 2020 year). The total number of projects inspected was 1015, and the inspection data included defect types, engineering levels, project costs, and construction progress (Table 1). Defects are classified into four types based on construction management (*X1*), work quality (*X2*), program (*X3*), and design (*X4*); the total classes are 499 defects. Among these, defects due to construction management include 113 classes for proprietors, supervisory sectors, and contractors. The defects due to work quality include 356 classes of safety, Strength I, and Strength II. Defects pertaining to the program include 10 classes. The defects due to design include 20 classes pertaining to maintenance, construction, security, and gender differences. Thus, the standardized forms have a total of 499 defective classes. *X1*, *X2*, *X3*, and *X4* accounted for 22.7%, 71.3%, 2%, and 4% of the percentage of defective classes, respectively (Figure 1).

Table 1. Construction inspection information and project feature content.

Decision Variables				Target Variables		
X1	***X2***	***X3***	***X4***	***Y1***	***Y2***	***Y3***
Construction management	Work quality	Program	Design	Engineering level	Project cost	Construction progress
Proprietors, supervisory sectors, and contractors (classes of 113 defects)	Safety, Strength I, and Strength II (classes of 356 defects)	Schedule, management, and project network diagramming management (classes of 10 defects)	Maintenance, construction, security, and gender differences (classes of 20 defects)	A, B, C, and D	P: publication (NT$1–50 mn) S: supervision (NT$50–200 mn) L: large procurement (NT$200 mn or more)	N: behind (under 50%) Y: ahead (over 50%)

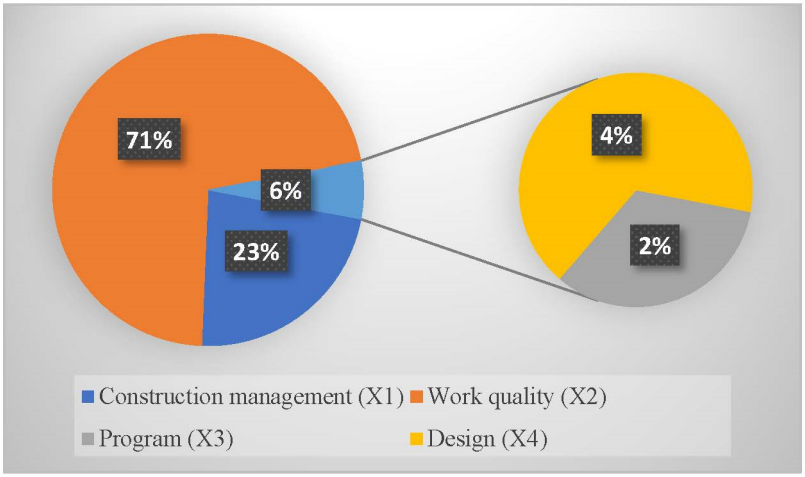

Figure 1. Proportion of defective classes in decision variables.

During the training of the classifier, the number of samples of different categories varies significantly (more than 20%), resulting in the machine learning algorithm mistakenly treating the samples of a few categories as tolerable errors and classifying all samples into the majority of categories to achieve a higher classification accuracy. Accordingly, the data imbalance should be addressed before constructing the model. The number of category samples in the training data can be balanced by processing it from two perspectives: the data and algorithm. Processing from the former perspective can improve the classification results of minority categories through sampling techniques such as undersampling, oversampling, and synthetic minority oversampling techniques (SMOTEs). Furthermore, the risk of random sampling can be reduced by using cluster-based sampling to divide the data into clusters and then identifying representative samples from the clusters [18]. Processing from the latter perspective involves adding a penalty item to the algorithm function, which increases the cost of misclassifying a certain category or multiple categories (such as cost-sensitive learning).

In particular, there are additional limitations on either the sampling method or the addition of penalized terms (penalized), which also tend to overfit the model. In the presence of data imbalance, the performance of classification models can be determined more objectively. The metrics used are usually precision and recall. Since the precision and recall of a good model are not too bad, neither are easily affected by data imbalance. Therefore, the harmonic mean (F1 score) of precision and recall was used as a measure of the imbalance classification problem. The F1 score is explained in detail in Section 4.3.

The engineering levels are regrouped using cluster analysis. Because of the present study, statistics from PWBMS on 1015 projects of construction inspection data found that six projects were Level I, 780 projects were Level II, 227 projects were Level III, and two projects were Level IV. The committee reviewed construction projects adopting predominantly Level II (collectively accounting for 76.8%), and scores were between 80~83. In comparison, cases of construction projects rated as Levels I, III, and IV accounted for only 23.1%. To reduce differences between numbers of samples and imbalanced data for all Levels (exceedingly few Level IV samples are available) and perform a comprehensive assessment, a cluster analysis of defect frequencies and the scores was conducted on the 1015 construction cases. Samples are partitioned into disjoint clusters based on their similarities or differences among feature variables. Cluster analysis avoided the sample size for each group from being extremely small and assured that defect features within groups were as similar as possible. The cases were reclassified into four groups (Levels A, B, C, and D), corresponding to Level I, Level II, Level III, and Level IV for construction inspection. The engineering level (*Y1*) were divided into four grades (A, B, C, and D). Cluster analysis was performed for these four levels by using two variables—defect frequency (18,246) and an inspection score of 1015 to reduce the number of samples sampled at different levels. The A, B, C, and D levels were 455 (44.8%), 157 (15.5%), 309 (30.4%), and 94 (9.3%) of all projects, respectively (Table 2).

Table 2. Statistics on the number of construction inspections and clustering.

Inspection Level	Score	Project Number	Clustering	Project Number
Level I	90–100	6 (0.6%)	A level	455 (44.8%)
Level II	80–89	780 (76.8%)	B level	157 (15.5%)
Level III	70–79	227 (22.4%)	C level	309 (30.4%)
Level IV	<70	2 (0.2%)	D level	94 (9.3%)

The project cost (*Y2*) was divided into three types—publication (P, NT$1–50 million), supervision (S, NT$50–200 million), and large procurement (L, NT$200 million or more). P, S, and L were 654 (64.4%), 138 (13.6%), and 223 (22.0%) of all projects, respectively. Construction progress (*Y3*) was divided based on the work progress into 50% or less and 50% or more, and these categories were 573 (56.5%) and 442 (43.5%) of all projects, respectively (Figure 2). The number of differences in the category data samples of the above

three target variables does not exceed 20%, which should satisfy the modeling requirements for the classification.

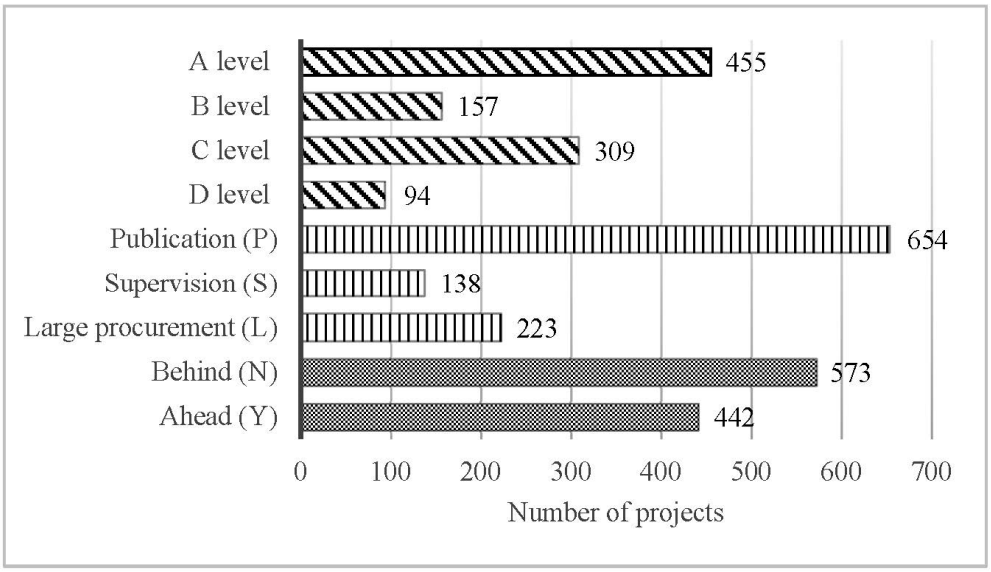

Figure 2. Number of project features of the target variables.

2.2. Application of Machine Learning in Engineering

The development of AI has evolved from "inference" advanced to "knowledge". Finally, AI has evolved having the ability to "learn". Therefore, while building a computer system, AI can intelligently recommend decisions by imitating human inference. Thus, it can prepare a computer with an accuracy similar to that of experts. ML involves the field of computer science and is a branch of AI. It is a means to realize AI, that is, ML has the ability to learn to solve problems. The following text presents a related application of ML in engineering.

A decision tree (DT) is an ML method commonly adopted for big data analysis with the main objective of formulating decision rules and conducting various classification works. The method is also applied in financial industries, manufacturing industries, and medical treatments for bank loan evaluations, examination of manufacturing defects, and disease diagnoses, respectively [19–21]. In the DT method, a tree structure pattern is utilized to interpret a series of decision problems and classification. DTs have been employed in the area of construction and management to predict and classify a widespread variety of attributes [22]. Some researchers have utilized the DT technique in construction for comparison with other ML methods [23,24].

The Bayesian network (BN) is a type of directed acyclic graph with directional and non-cyclical conditions based on conditional probability, and it can combine uncertainties in specific fields into a model [25]. A BN mainly relies on the advantages of a graphical model to effectively infer the uncertainty due to a large number of variables to determine the modeling domain. Recently, many studies have applied BNs to solve problems related to engineering structures and structure strength. For example, Straub and Kiureghian [26] combined a BN and the structural reliability method to create a new computation framework for risk analysis of infrastructure and engineering structures. Ma et al. [27] proposed a method involving BN to predict bridge residual strength. BN is also used for conducting qualitative and quantitative estimations of the affects of reinforced concrete structures [28], and for executing classification in maintenance inspections of road structures [29].

In an ANN, biological neurons are simulated to acquire information from other artificial neurons or the external environment. The neurons are learned using the network structure and different algorithms for performing estimation, prediction, and decision making. Using neurons, the network learns the patterns of the dataset, establishing the ability

to accurately classify new models, and carrying out predictions [30]. ANN applications in construction include the evaluation of project costs [31–33], construction safety [34,35], construction risk assessment [36,37], and diagnosis of related defects [38].

SVM uses a supervised learning algorithm and is a learning method for addressing classification or regression problems. An SVM obtains support vectors from specific training data by addressing mathematical problems of quadratic programming [39]. An SVM is frequently applied for spatial feature identification of image data and in engineering structures such as detection or classification. Li et al. [40] implemented bridge crack recognition to use the greedy search-based support vector machine. Hadjidemetriou et al. [41] used the SVM to automate pavement patch detection and quantification. Liu et al. [42] proposed a technique for fire damage identification of reinforced concrete beams with the SVM. Chen et al. [43] exploited an SVM-based evaluation approach and rust recognition for steel bridges.

Additionally, ML can extract complicated connected modes of parameters potential in large amounts of data and is an adequate method for building a classification model of the concrete defect [44]. Okazaki et al. [11] stated the importance of employing ML to defect prediction for infrastructure. In the past 20 years, some scholars have proposed multiple hybrid ML algorithms to analyze construction defects. These algorithms include DT, BN, ANN, SVM, cluster analysis (CA), association rules (AR), genetic algorithm (GA), and fuzzy logic (Table 3).

Table 3. ML algorithm analysis using project defect statistics and comparisons.

Author	Algorithm	Description
Sinha and Fieguth [7]	ANN and fuzzy logic	ANN and fuzzy logic are proposed for the detection of defects by using features from underground pipe images.
Cheng et al. [8]	GA and AR	Using Genetic Algorithm (GA)-based association rules (AR) enhanced construction management by causation analysis and defect prediction.
Elmasry et al. [9]	BN	The BN developed a defect-based deterioration employing the likelihood of occurrences from the sewer pipelines.
Lin and Fan [10]	AR, CA and fuzzy logic	Combining association rules (AR), cluster analysis (CA), and fuzzy logic, they mine the causal relations between inspection grades and construction defects.
Lee et al. [13]	AR	AR is used to find the between defect causality, and social network to explore indirect causality among defects.
Chae and Abraham [45]	ANN and fuzzy logic	The proposed multiple ANN and fuzzy logic recognize the various conditions of defects from sanitary pipelines.
Cheng and Leu [46]	CA	Integrating-based clustering and affinity diagram (KJ method) to classify bridge construction defects.
Gui et al. [47]	GA and SVM	The genetic algorithm based SVM had for the large-scale civil engineering structures to detect the damage/defects.
Lin and Fan [48]	DT	DT were developed for classification rules of defects to compare the performance of CART, CHAID, and QUEST.
Fan [49]	AR and BN	The AR and BN (hybrid ML) approach was adopted to find relationships and probability in construction defect data and assess the risks of defects.

3. Research Methodology

The utilization of ML involves a computer system calculation that addresses problems intelligently by imitating the thought process of the human brain [50] in solving prediction or classification issues [51]. In ML, many algorithms and technologies, such as DT, ANN, and SVM, have been used to execute prediction tasks and classification models [52,53]. Classification and regression are used to acquire a group of models by training information, and the model can be applied to predict the group category of unclassified data. Both models build a system of feature (input) and label (output) relationships. The goal is to establish a model that can be applied to predict unknown data with similar characteristics. Moreover, the classification and regression outputs are discrete and continuous values, respectively. In the former, the data are assumed to be correctly labeled, whereas in the latter, the data class labels are derived from random variables [54].

The ML algorithm intelligently learns and acquires rules from data, then analyzes these rules to classify unknown data. The aim of this study is to analyze the hidden pattern of the defect information gathered by the committee using the construction inspection data in public projects. Moreover, ML was utilized to identify potentially valuable features related to target variables and defects and to assess the benefits of the classification.

3.1. Decision Trees (DT)

DT sets target variables and selects branches to present as a hierarchical structure to explore rules. A pruned DT can discover the hierarchical relationship between decision variables and target variables, and use it for prediction. A DT is a supervised ML algorithm that can be employed as an effective technology for multivariate analysis [55]. DT algorithms including C5.0, classification and regression tree (CART), Chi-squared automatic interaction detector (CHAID), and quick unbiased efficient statistical tree (QUEST) are commonly applied in academia and the industry [56].

3.1.1. CART Algorithm

Breiman et al. [57] proposed the CART algorithm is a binary splitting technology. After training the CART, pruning is implemented, and the error proportion is adopted as the base for pruning. The CART with the least number of tree layers provides the most valid classification and applies to target variables expressing categorical and continuous data. Accordingly, the target variables represent categorical and continuous data, then classification and regression trees may be employed, respectively. The splitting condition in the CART algorithm is determined based on the Gini index. The object of the Gini index is to determine the largest number of classifications from the dataset in other categories at different nodes. The smaller the Gini index, the more uneven the category allocation of the data is. This implies that if the purity of category in the subset produced using the splitting point is higher, then the capability to discriminate among different categories is better. If a dataset (S) includes m data, F_i is the relative frequency that the data of category i occurs in S. The equation of the Gini index is given as follows.

$$\text{Gini}(S) = 1 - \sum_{i=1}^{m} Fi^{2} \quad (1)$$

3.1.2. CHAID Algorithm

Kass [58] proposed CHAID as an implemented Chi-squared test to determine the splitting condition. Moreover, a probability value determines whether the splitting operation in the CHAID has to be continued to evaluate possible predictive variables. In the algorithm, the significant differences among the categories of dependent variables are tested with all variables. When the insignificant categories are merged into a homogeneous class, and the remaining categories are repeatedly examined until the differences become insignificant. The CHAID is used to compute the feature branches, and they can be separated into merge and split. In the first step, each variable is regarded as a different group, and the value of

the merger is set. In each operation procedure, the two branches of the tree are pairwise comparisons primarily to define whether *p*-value significantly differs. If the *p*-value is greater than the merger value, and the two branches are merged into one branch, which represents that the required significance level is not achieved; then, the examination is repeated. The procedure is continued until all the outcomes obtained after the pairwise branch examination are significant. The second step, the value of the split is set, and all branches containing more than two categories are examined. If *p*-value is less than the split value, the variables of different categories are split into different branches, which represents that branch being significant.

3.1.3. QUEST and C5.0 Algorithm

Loh and Shih [59] proposed the QUEST algorithm, which is used for classifying tree structures. The splitting rule in this algorithm assumes that the target variable is continuous. This algorithm has a faster computation speed than the other ML based on DT and can also prevent the bias present in those algorithms. The QUEST algorithm is more suitable for mult-category variables; however, it can only calculate binary attribute data. Quinlan [60] proposed that the C4.5 is incapable of processing continuous attribute data. C5.0 was developed to overcome this drawback. It adopts information gain as the standard for determining the variables of branches and uses cross-validation and boosting training data for faster and more accurate DT analysis. The C5.0 algorithm evaluates the amount of information under different conditions to find the characteristic that can obtain the maximum gain based on the information frequency of categories. As in the DT branch, the overall information involving all categories is presented in Equation (2). The frequency of occurrence in the category can be defined as F_i, and the information of the category is $-\log_2 F_i$. The above four DT algorithms are summarized in Table 4 [48].

$$\text{Info}(S) = -\sum_{i=1}^{j} Fi \times \log 2(Fi) \tag{2}$$

Table 4. Description of decision tree algorithms.

Attribute	CART	CHAID	QUEST	C5.0
Variable type	Continuous or category	Category	Category	Continuous or category
Number of branches	Two	More than two	Two	Continuous: more than two; Category: two
Branching variable	Single or multiple variables	Single variable	Single or multiple variables	Single variable
Splitting rule	Gain index	Chi-square test	F/Chi-square test	Information gain
Prior probability of classification	Yes	No	Yes	No
Tree pruning	Test sample or cross-validation	Stopping rules	Test sample or cross-validation	Simultaneous branching and pruning

3.2. Bayesian Network (BN)

BN is a probability model, and it adopts a graphical pattern to describe the relationship between variables that graphically displays statistical influence by utilizing the causality between variables for conducting predictions. The aim of using a BN is to analyze the probability of an uncertain event in a decision problem by using a set of random variables and to determine the influence relationship between the variables. BN can be modified

at any time based on new information or evidence and then pushes out the posterior probability of uncertain events [49].

A BN, that is, $D = < M, N >$, a set of conditional probability distribution (CPD) elements and consists of a network structure. In Part I, M is a set of dependent or conditional independent relationships in the model, which is directed acyclic and describes the network structure built by a set of variables $X = \{X_1, X_2, X_3, \ldots, X_n\}$. These variables are represented by nodes, and the relationship is represented by a link. In Part II, N is a set of CPD elements connected with variables. Parents (x_i) represents the parent node of X_i in M, and P(X_i | parents(x_i)) represents the CPD of node X_i under the parent node (parents(x_i)). The joint probability distribution P(X) is combined with M and N. The equation is as follows.

$$\begin{array}{l} P(X) = P(X_1, X_2, X_3, \cdots X_n) \\ = \prod_{i=1}^{n} P(Xi \mid x_1, x_2, x_3, \cdots, x_n) \\ = \prod_{i=1}^{n} P(Xi \mid parents(xi)) \end{array} \quad (3)$$

In summary, BN displays a set of random variables, and the n sets of CPD models are obtained using graphs of links and nodes. A model with n CPDs is used to present the relationship and strength between the variables. A link between the variables represents an interaction between the events, and a node represents a variable (such as a latent variable, a variable of an observed value, or an unknown parameter). When there is no connection between the nodes, it implies conditional independence.

3.3. Artificial Neural Network (ANN)

ANN is an information operation system that imitates biological neuronal networks, can receive information from other neurons or external environments, and can solve complex problems. ANN uses different learning algorithms to ensure that it outputs the desired result and is trained by a network structure. ANNs are constituted of many artificial neurons, wherein the output of each neuron is used as an input for other neurons. ANNs typically use a set of data to develop a model to predict, classify, and estimate. The equation is as follows.

$$Y_i = f(\sum W_{ij} X_i - \theta_j) \quad (4)$$

Y_i and X_i are the output and input of the neuron signals, respectively, and f is the activation function. The purpose of the activation function is to multiply the values input by other neurons with the weight and add all values to convert the output values of the neuron. W_{ij} (weight) is the connection strength of the ANN, and θ_j is the threshold of the neuron model. An ANN comprises multiple nerve cells. Each of the cell has a weight value W_{ij} for indicating the influence intensity of the i-th input on the j-th output. When all input values are multiplied by the weight, the total value is greater than the threshold (θ_j). The total value is converted to an output value through the activation function, and it is passed to the next neuron. For ANNs, multi-hidden layers in which deep learning is used to gradually summarize higher-level features from input data have been proposed [61,62].

3.4. Support Vector Machines (SVM)

An SVM is an ML algorithm proposed by Vapnik [63] that is a type of classification model and is based on the statistical learning method. The basic theory of an SVM is to identify the optimal hyperplane of the boundary in high-dimensional space to classify binary categories and obtain a minimum misclassification rate. The optimal hyperplane is segregated based on maximizing the margin between categories in the features and ensuring reasonable generalization capacity of the result [64]. SVM is trained using existential data, and the analyzed data are used to select several features (support vectors) to represent the overall data. A small number of extreme values are removed in advance, and then the selected support vectors are packaged into models.

The major steps in this study are shown in Figure 3. In this work, the first stage is to select target variables and decision variables from PWBMS and evaluate the classification

models of seven ML algorithms. The second stage applies cross-validation to test results and develops an optimal model of classification for project features.

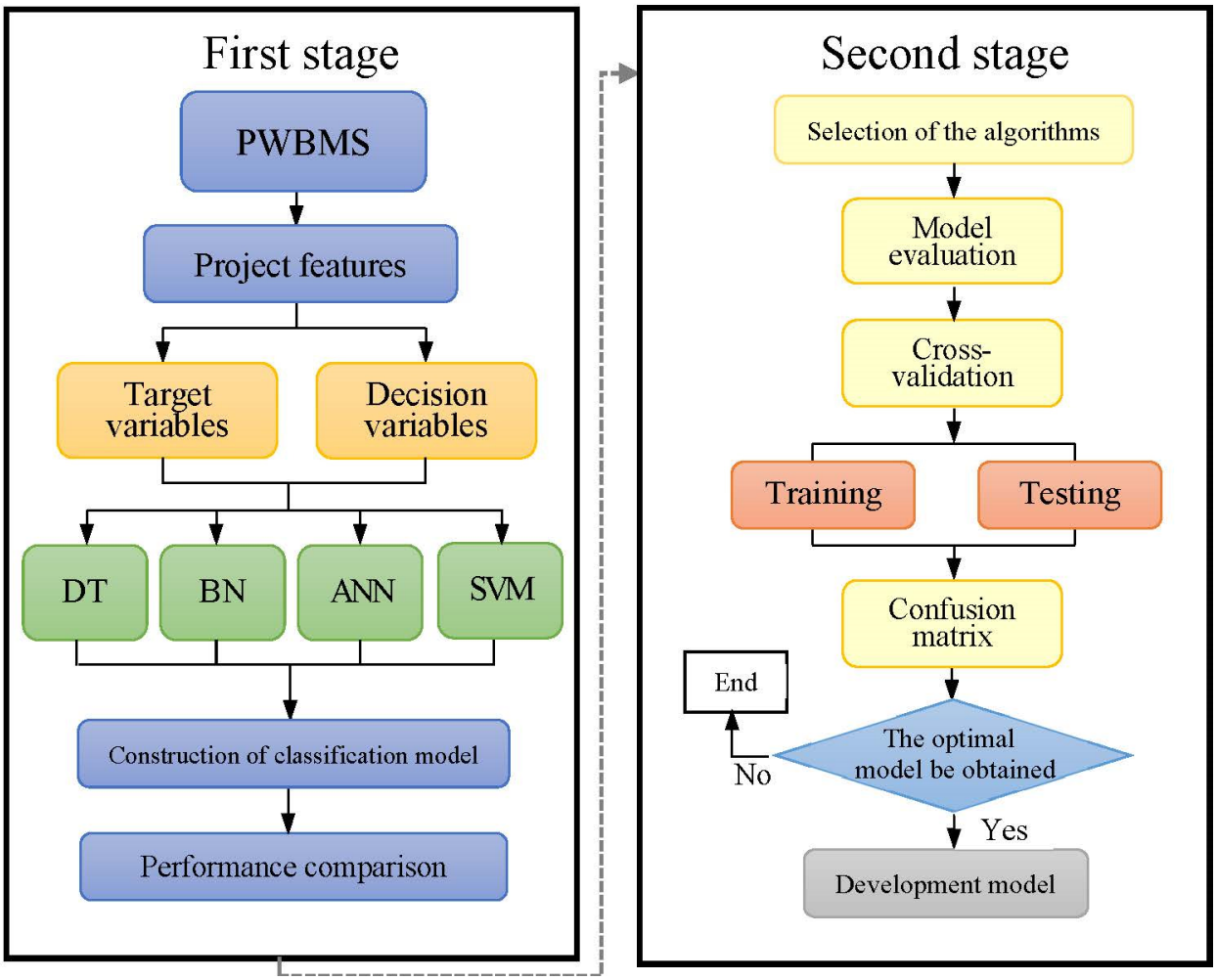

Figure 3. Process steps for the study.

4. Analytical Results

ML constitutes developing algorithms or models for predicting results by learning from the data features; it focuses more on predictive accuracy and computational efficiency [54]. ML has the capability to identify structures and patterns hidden in massive datasets, without assuming a predetermined function as a model [11]. ML algorithms are divided into two learning models: unsupervised and supervised learning. Most ML algorithms are supervised; labels are used to train a model using known answers (input), after which the model can predict output for new input data.

4.1. Cross-Validation

In the supervised ML algorithm, a machine learns from the training data to perform better models in the testing data. However, deviations (errors) exist between predicted outputs of the training and testing. The errors of the training and testing groups are known as the "training error" and "generalization error", respectively. The preferred supervised ML demonstrates a smaller generalization error. However, the property of the new data cannot be expected, and actually executing can only minimize training errors. If the training model is highly complex, then overfitting is caused. Moreover, if the model is very simple, then underfitting is caused. Supervised ML encounters the problems of overfitting and underfitting, minimizing the difference between the two. Although it is possible to conduct

more iterations to further reduce the error, long training with very few errors will result in an overtraining problem and, as a result, the models will memorize the unique training patterns of the samples presented but be unable to generalize them [65].

To make the model meet the training group after a certain amount of learning and training, it can better adapt to the new sample and measure the learning level of the model. The performance of the model must be validated while providing the best choice among multiple models. Thus, cross-validation avoids reliance on particular training and testing groups, and it is a measurement for evaluating reliability and accuracy. Cross-validation is one of the most important concepts in any type of data modeling. It tries to leave a sample set, does not train the model on this sample set, and tests the model on that sample set before finalizing the model.

A k-fold of cross-validation randomly divides data into two groups containing training and testing groups. Training data are divided into k subsamples, and a single subsample is retained as the data for the test model. The remaining $k-1$ samples are used for repeating training k times. Each subsample is tested once, and the average k is the obtained accuracy of the results. Because the statistical effect has not improved much in a larger number of cases, the value of k should be between 5 and 10 [66]. Therefore, this study used tenfold cross-validation to randomly divide construction inspection data into 10 groups. One group was used for testing, and the remaining nine groups were used for model training (Figure 4). The advantage of this method is that the subsamples are generated simultaneously, the subsamples are trained and tested repeatedly, and the results are verified once. The accuracy of each model was evaluated using the average prediction results of the ten groups. This method can reduce the bias during the model evaluation and can overcome the unevenness in data caused by sampling only once, and the classification model prediction is not sufficiently accurate. While estimating the performance of each model and comparing two or more accuracy prediction methods, the errors caused by random sampling can be reduced to evaluate the benefits of various feature classifications and the reliability of the model.

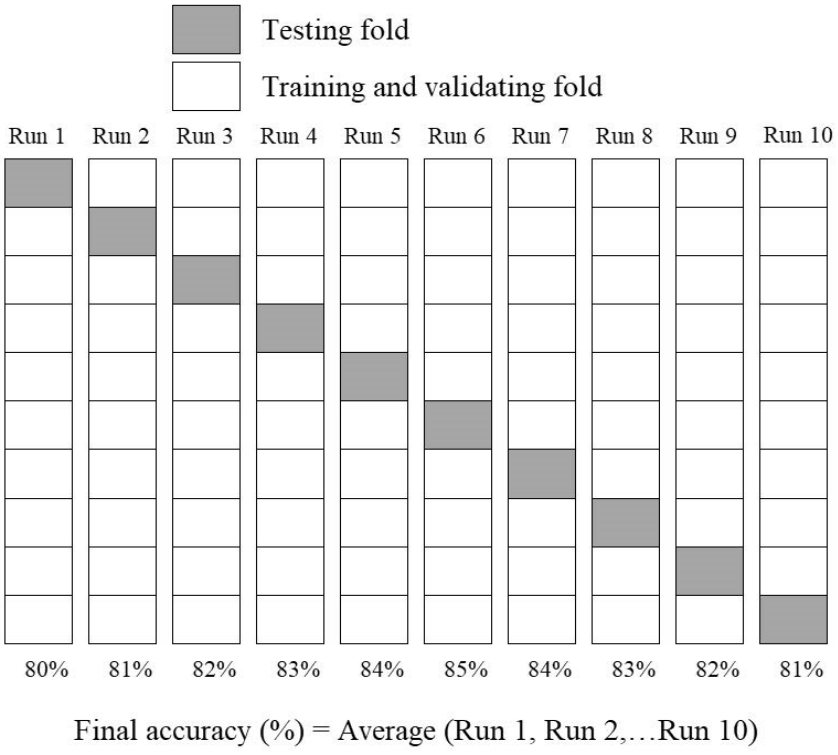

Figure 4. Tenfold cross-validation process.

Classification involves assigning samples to a predefined category, and the target variables are labeled, representing the category membership of input data. A ML algorithm identifies a classifier set of feature variables and exploits a model to predict category membership of new data with unknown labels, based on the classifier set identified [54]. This study builds a classification model in which construction features are outputted as responses, and defect types are inputted as predictors. This research applies to the following seven types of classification algorithms in the SPSS modeler software: CART, CHAID, QUEST, C5.0, BN, ANN, and SVM algorithms. To construct seven classification models, the attributes of 1015 cases of construction inspection were analyzed. The selected decision variables were *X1–X4* (defect types), and the target variables were *Y1–Y3* (project features).

The most major consideration in the model is the impartial evaluation of its performance. To achieve this goal, the data of construction inspection is divided into training groups and testing groups. Training and testing models is an important process for implementing supervised ML algorithms. Therefore, seven models of tenfold cross-validation were used in this research, and the average accuracy of classification models are summarized in Table 5. The ANN for engineering level (*Y1*) generated the best prediction results with an accuracy of 88.24%, in the 1015 projects of construction inspection data. The SVM obtained the best results for project cost (*Y2*) and construction progress (*Y3*) with accuracies of 78.91% and 76.18%, respectively.

Table 5. Accuracy of the classification models for the target variables.

Target Variable	The Best of Testing Set (%)	The Average Accuracy of Tenfold (%)						
		CART	CHAID	QUEST	C5.0	BN	ANN	SVM
Engineering level (Y1)	84.87 (ANN)	72.31	70.37	66.03	76.47	58.52	88.24	80.74
Project cost (Y2)	73.75 (SVM)	68.42	69.49	68.81	71.45	61.2	69.59	78.91
Construction progress (Y3)	71.46 (SVM)	58.54	61.9	59.04	68.49	56.13	58.47	76.18

In addition, this study used the SPSS modeler automatic classifier to construct a predictive evaluation model and used construction inspection data to train the model. The accuracy of the model was evaluated using the testing set and was compared with the results of the binary data. Because only a portion of the data was used to derive the classifier, the evaluation results are not necessarily optimized. Therefore, it is necessary to compare the evaluation results with the cross-validation results.

Furthermore, the target variables contained the engineering level (*Y1*), project cost (*Y2*), and construction progress (*Y3*) in the testing group. The ANN can accurately classify engineering level (*Y1*), and the SVM can gain accurate results of project cost (*Y2*) and construction progress (*Y3*). In this study, the results of the models were consistent with the cross-validation (Table 5). In the engineering level, the ANN correctly classified the numbers of A, B, C, and D as 205, 68, 132, and 38, respectively (Figure 5). The accuracy of the testing set was 84.87%. In project cost, the SVM correctly classified the numbers of P, S, and L as 339, 18, and 28, respectively, and the accuracy of the testing set was 73.75% (Figure 6). In construction progress, the numbers of correct classification of the SVM were 283 (N represents less than 50%) and 90 (Y represents more than 50%), and the accuracy of the testing set was 71.46% (Figure 7). In general, the SVM demonstrates better classification accuracy for the three target variables (project features), followed by C5.0; however, the ANN has the most illustrious classification for engineering level.

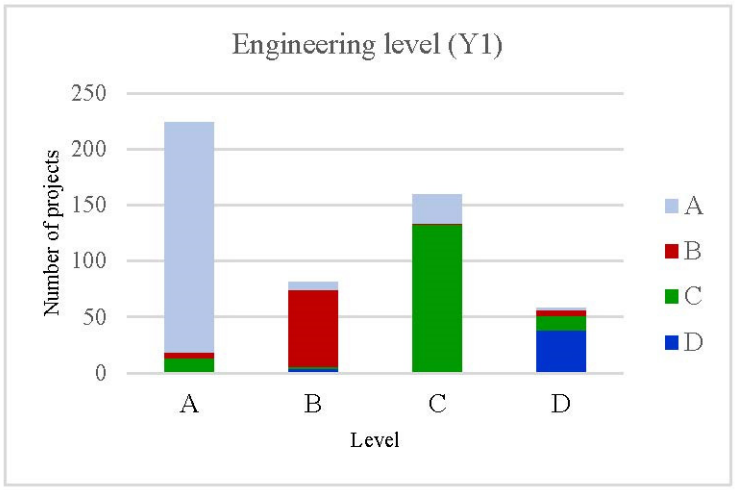

Figure 5. The numbers of ANN classification for engineering level.

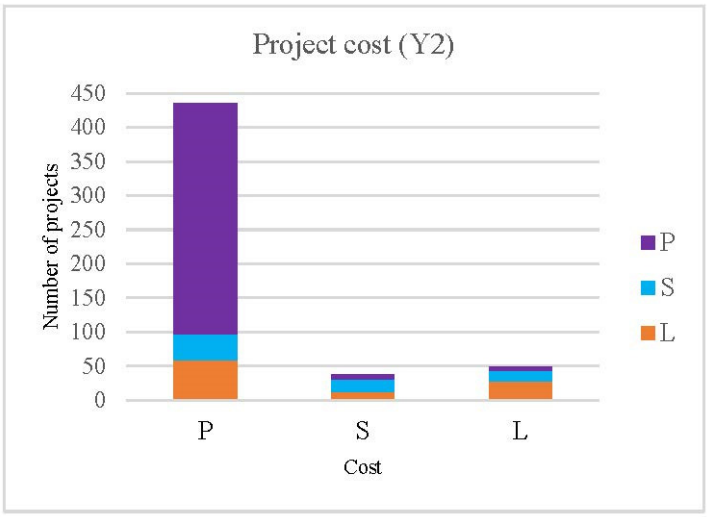

Figure 6. The numbers of SVM classification for project cost.

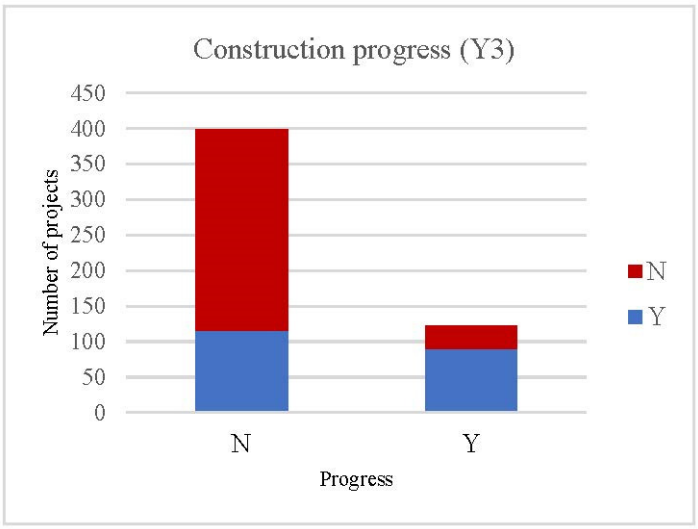

Figure 7. The numbers of SVM classification for construction progress.

4.2. Importance Value of Defect

In this research, importance values were applied to indicate the relative importance of features. A higher importance value of a feature represents powerful classification ability. Feature importance was assessed by calculating the total decrease in the impurity of each feature. For the given algorithm, the sum of importance values of all selected features should be 100% or 1 [67]. In this research, the importance values of the features were between zero and one, and the total of the relative importance values of all defects should be one.

Figure 8 presents the importance value of the C5.0, ANN, and SVM models for defects in the engineering level (*Y1*) classification. It is obvious that the most important defect identified by the three models was "substandard concrete pouring or ramming (W2)", which implies that W2 has the highest classification capability of all defects. W2 was followed by "debris on concrete surface (W5)" and "failure to log the construction journal (Q75)". However, the C5.0 algorithm excludes Q75, thus indicating that Q75 has nonsignificant predictive ability in the C5.0 algorithm. Therefore, unimportant input features are preferentially deleted.

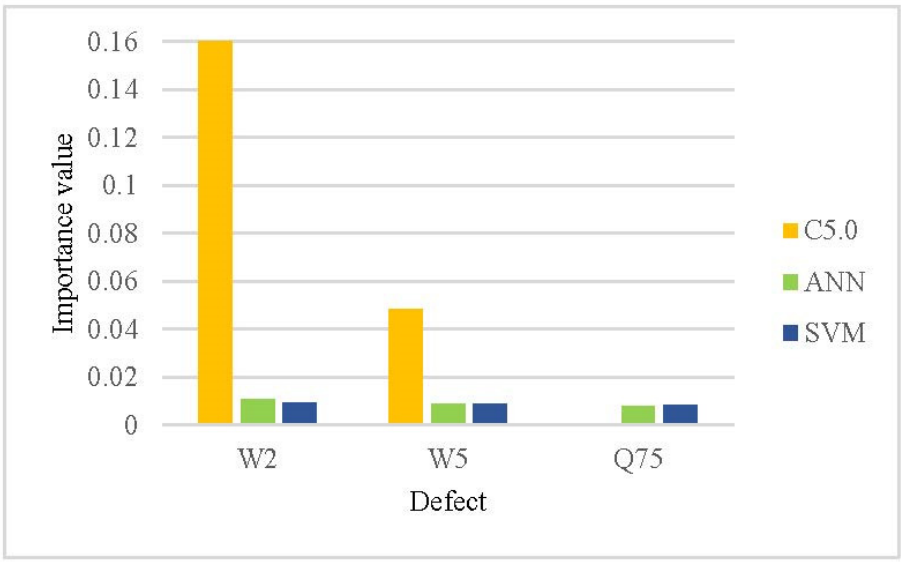

Figure 8. Importance values of defects in C5.0, ANN, and SVM for engineering level.

Both ANN and SVM models selected W2, W5, and Q75 as relatively important defects, and the important values of the three defects were similar. The C5.0 algorithm is relatively more important for W2. If the target variable is changed to "project cost (*Y2*)" and "construction progress (*Y3*)" to perform another round of testing, cross-validation results reveal that the SVM can produce the best prediction accuracy. The critical defect of the project cost is "failure to implement a quality control checklist (Q76)" and that of the construction progress is "debris on concrete surface (W5)".

4.3. Confusion Matrix

The model of supervised ML yields different results for different algorithms, and the prediction performance can be evaluated from the two parts of classification and regression. (i) Classification: the classification model should be verified against the results of the test set data and evaluated using a confusion matrix (output value is discrete). (ii) Regression: the extraction of classification rules vary based on the problem to be solved, and the difference in rule interpretation varies due to the environment. Thus, after the objective evaluation, experts and scholars select the most suitable model based on the background of the problem (output value is continuous). If the variable of the predicted classification label has only two values of zero and one, it is a binary classification. If the label variable of

the predicted classification has more than two values, it is a multiclass classification. In this study, construction progress (*Y3*) is a binary classification, and the engineering level (*Y1*) and the project cost (*Y2*) are multiclass classifications.

The confusion matrix is an important tool for evaluating the classification model. There are two categories of a model, positive and negative, and the prediction or classification result is consistent with the actual category of the data, which is known as "True". If the inconsistency is known as "False", a discriminating error is caused, assuming that the category is actually positive, but it is classified as negative (false negative, FN) and the category is actually negative, but it is classified as positive (false positive, FP). If the category is actually positive, the classification is also positive (true positive, TP), and the category is actually negative, and the classification is also negative (true negative, TN). Thus, based on the classification result presented in Table 6, and the accuracy of the model can be calculated as presented in Equation (5).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (5)$$

Table 6. Confusion matrix of the classification model.

Actual	Predicated	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

For example, the ANN correctly predicts that the value of engineering level A, B, C, and D, that is, 425, 135, 302, and 84, respectively, when divided by the total number of 1015. Thus, the accuracy is 93.20% (Table 7). ANN can process different types of original data, learn through data training, improve the accuracy of prediction, and establish a classification model. Compared with other classifiers, the disadvantage is that the ANN does not easily interpret the relationship between input and output during the operation. Moreover, the SVM prediction accuracies of the project costs and construction progress, which were 85.32% (Table 8) and 79.01% (Table 9), respectively, were used in this study. The evaluation results of the seven classification models are presented in Table 10.

Table 7. Confusion matrix and accuracy of ANN for engineering level (Y1).

ANN	Predicated				Total	Recall (%)	Accuracy (%)
Actual	**A**	**B**	**C**	**D**			
A	425	5	22	3	455	93.41	93.20
B	7	135	11	4	157	85.99	
C	2	1	302	4	309	97.73	
D	0	5	5	84	94	89.36	
Precision (%)	97.93	92.47	88.82	88.42	n = 1015	Mean = 91.62	

Table 8. Confusion matrix and accuracy of SVM for project cost (Y2).

SVM	Predicated			Total	Recall (%)	Accuracy (%)
Actual	**P**	**S**	**L**			
P	66	20	52	138	47.83	85.32
S	14	166	43	223	74.44	
L	6	14	634	654	96.94	
Precision (%)	76.74	83.00	86.97	n = 1015	Mean = 73.07	

Table 9. Confusion matrix and accuracy of SVM for construction progress (Y3).

SVM	Predicated		Total	Recall (%)	Accuracy (%)
Actual	N	Y			
N	549	24	573	95.81	79.01
Y	189	253	442	57.24	
Precision (%)	74.39	91.34	n = 1015	Mean = 76.53	

Table 10. Assessment results of algorithm on three target variables.

ML Algorithm		CART	CHAID	QUEST	C5.0	BN	ANN	SVM
Engineering level (%)	Accuracy	79.01	77.83	67.39	78.13	62.96	93.20	82.17
	Precision	79.54	76.8	65.13	75.57	87.0	91.91	92.01
	Recall	74.17	78.05	68.19	73.79	90.12	91.62	75.35
	F1	76.76	77.42	66.62	74.67	88.53	91.76	82.85
	BEP	80.0	79.30	64.60	74.0	87.20	90.80	95.30
Project cost (%)	Accuracy	76.16	78.82	71.72	73.89	65.02	80.89	85.32
	Precision	76.14	73.94	69.09	62.35	53.29	91.81	82.24
	Recall	58.18	64.39	48.93	56.97	53.98	64.11	73.07
	F1	65.96	68.84	57.29	59.54	53.63	75.50	77.38
	BEP	82.20	73.80	71.0	59.0	47.0	87.80	83.90
Construction progress (%)	Accuracy	62.96	64.14	58.72	70.25	58.62	75.37	79.01
	Precision	64.98	64.92	63.17	69.97	58.03	74.95	82.86
	Recall	58.97	60.89	53.12	70.26	58.07	74.9	76.53
	F1	61.83	62.84	57.71	70.11	58.05	74.92	79.57
	BEP	67.80	67.0	64.30	71.20	58.60	75.90	85.10

The performance of a classification model is expressed in terms of accurate estimates, particularly when the numbers of a certain category are relatively small and demand more attention. Therefore, only accuracy is employed; then, other categories with higher number of category rates are prioritized. However, extremely small numbers of categories may indicate valuable information, which can be used as an evaluation criterion by using precision and recall. Precision expresses the number of all prediction categories that actually belong to the category, as given in Equation (6). Recall refers to the fact that the actual result of a certain category correctly predicts the rates, as presented in Equation (7).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

If precision is low, recall is typically higher and vice versa, and both are a set of opposite measurements. The precision–recall (P–R) curve is acquired by plotting the precision and recall along the vertical and horizontal axes, respectively; according to the classification result of the ML classifier, the two sequences above are executed. The P–R curve visually displays the precision and recall of the ML classifier (Figures 9–11). If the P–R curve of the classifier is covered by the other classifier (completely contained), it can be judged that the classification performance of large curves is better than the smaller; however, the P–R curves of the two intersect, and it is difficult to judge the performance. Therefore, it can be measured by using break even point (BEP). When the precision and recall values are the same, the point of intersection in the P–R curve is BEP, and the recall value of BEP is adopted. In this research, the ML algorithms with the highest BEP values for engineering level, project cost, and construction progress were SVM (95.3%), ANN (87.8%), and SVM (85.1%), respectively (Table 10).

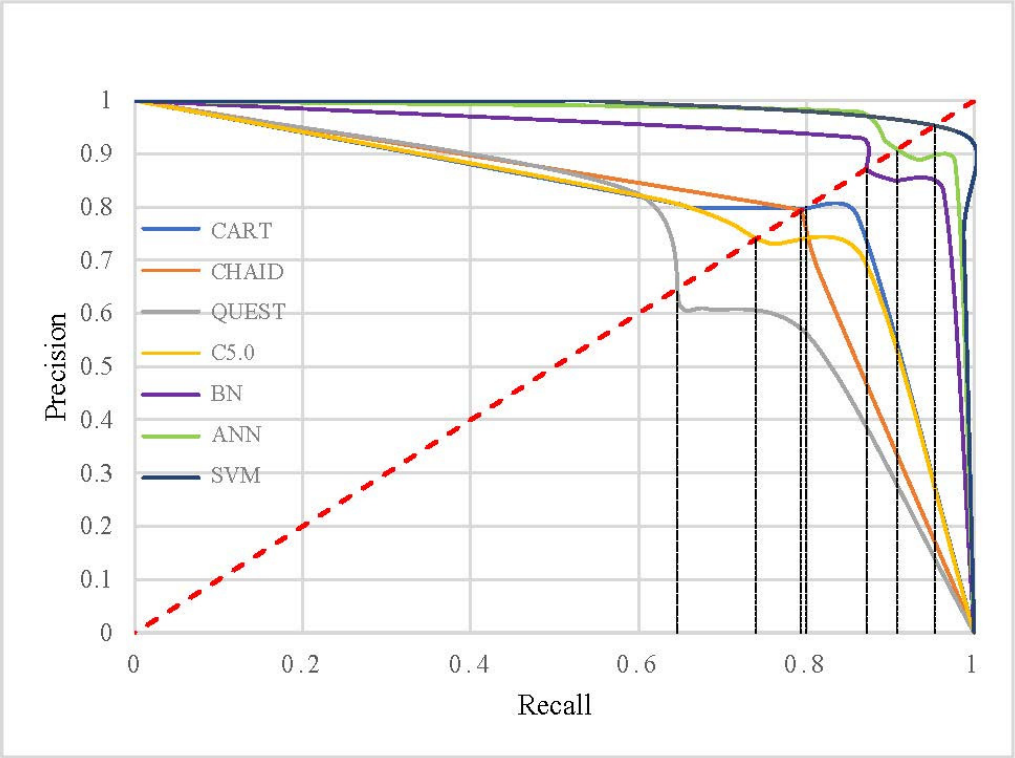

Figure 9. P–R curve and BEP of the ML algorithm for engineering level.

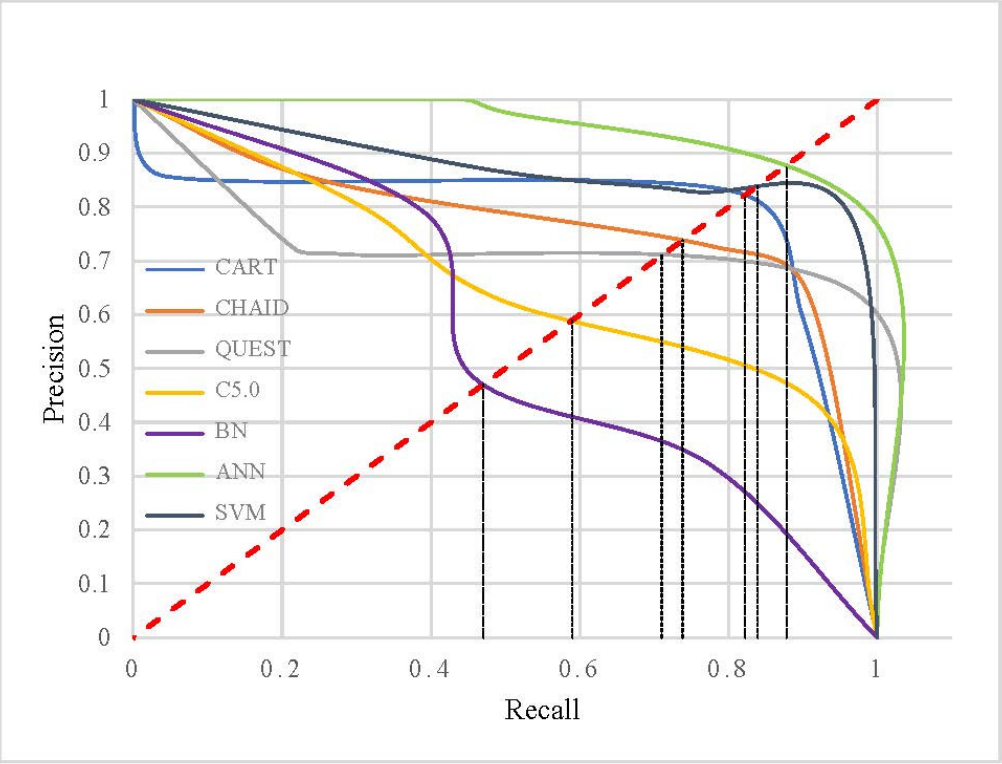

Figure 10. P–R curve and BEP of the ML algorithm for project cost.

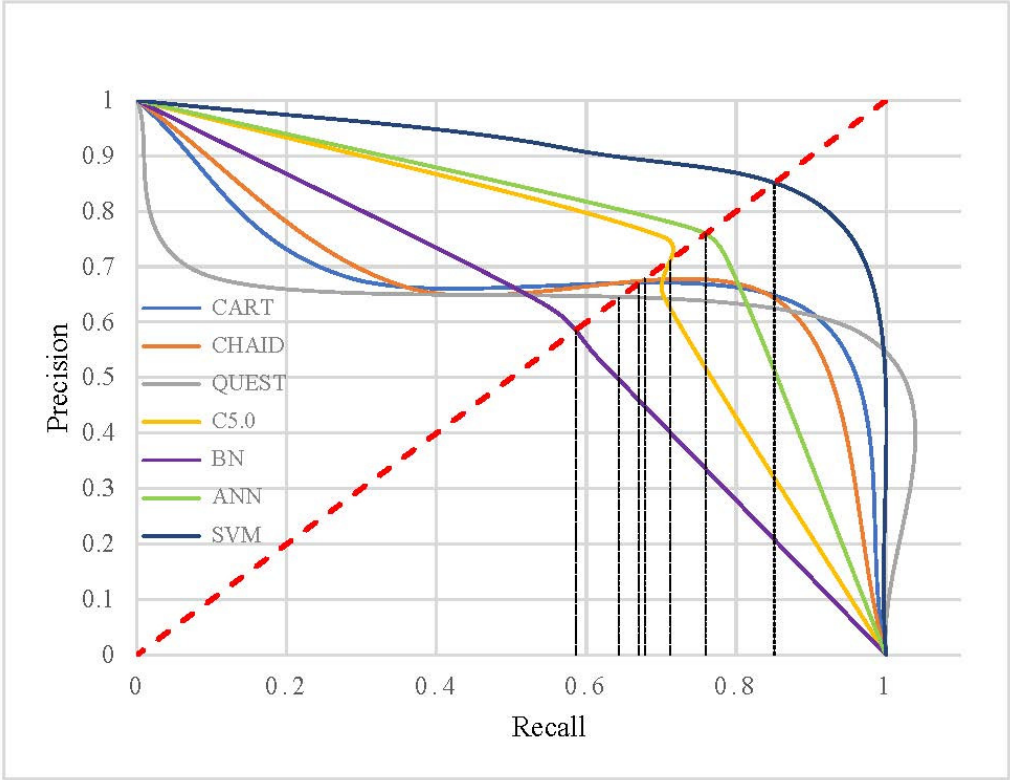

Figure 11. P–R curve and BEP of the ML algorithm for construction progress.

The F1 score has been widely employed for imbalanced data classification, which is an assessment metric determined by combining precision and recall [68]. In the training classification, the expectation is that the precision is as high as possible and also that the recall is as high as possible. In some examples, the two indexes are negatively correlated. The F1 of the model is high when both precision and recall scores are high. In general, the mean calculation treats each value equally, but the harmonic mean gives higher weight to smaller values. Therefore, the adoption of the F1 score harmonizes the two (as shown in Equation (8), and the higher the F1 score, the higher the performance of the model. In this research, the highest F1 scores in the engineering level, project cost, and construction progress were obtained for the ANN (0.918), SVM (0.774), and SVM (0.796), respectively (Table 10).

$$\mathrm{F1} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{8}$$

DT is the best in generating and judging rules, but the model is poor in predicting the ability to process continuous values, and too many types of features have a great impact on the results and their performance. The ANN model can easily construct nonlinear input variable models, which is more flexible, but the relationship between the output variables and the input variables of the process is not easy to explain. In spite of the ascendancy of SVM over other machine learning techniques, it suffers from optimization problems, such as the trade-off minimization error and between maximization margin. SVM may face issues with selecting items for classification and suffer from different performances in the classification, relying on the kernel functions that are used. Thus, for the practical application of such multi-class classification results, it is also necessary to select the general criteria that are suitable for the algorithm.

In the past, research on defects mostly used traditional statistical methods or multivariate techniques. It was impossible to effectively analyze multi-dimensional data features and dynamic projects. In particular, it is difficult to perform analysis using traditional data processing techniques for big data, while machine learning is utilized to design models to

learn tendency, so as to focus on predictions based on known features learned from the training datasets. Therefore, importing ML can analyze the features of defects from a large number of projects and then implement effective construction management. So far, there has not been a sufficient model to mine meaningful information from a large database of defects. This research can establish more prediction applications in construction defects and contribution through comprehensive and accurate model results.

Each ML method has its own advantages and limitations, and the characteristics of its application define which method is most suitable. Therefore, it has become important to shift the various algorithmic options from exploratory to targeted and reasonable implementations, since different ML algorithms can produce different levels of accuracy and performance depending on the application [69]. In addition, different machine learning methods are adopted according to the collected information, which is also related to the type of issue to be addressed: determining the nature of the problem, testing machine learning algorithms, and evaluating the suitability of these algorithms for a given problem. Wolpert and Macready [70] have stated that it is impossible to adopt unique optimal methods, and the best technique always relies on the character of the issue. The selection of the appropriate ML algorithms depends on characteristics and the amount of training data in the dataset, and the evaluated features being considered [71]. There is no one algorithm that has universal superiority for all problems; instead, there must be an algorithm that performs best in solving a certain type of problem. Each machine learning method has its own data that are suitable for processing, and there is no absolute perfect method; especially, when facing complex problems and a large amount of data, different algorithms are usually needed to effectively overcome them. The novelty of this study is to discover the correlation of project features from massive construction data through ML algorithms and to construct an optimized classification model based on classifier performance.

5. Conclusions

In this research, supervised ML algorithms (DT, BN, ANN, and SVM) were selected to classify projects and, according to the features of the construction data, were adopted. Seven classification models were built, and the classification accuracy and performance were evaluated. The cross-validation results demonstrated that the classification models adopting SVM, ANN, and C5.0 revealed higher classification efficiency and more reliability than other ML algorithms. The confusion matrix evaluation results revealed that ANN yielded the highest classification accuracy for engineering level (*Y1*: 93.20%). Moreover, SVM presented the most favorable classification result for project cost (*Y2*) and construction progress (*Y3*) with accuracies of 85.32% and 79.01%, respectively. In general, the SVM had better classification performances for the three target variables (project features). The most important defect of the ANN for the engineering level (*Y1*) classification model was "substandard concrete pouring or ramming (W2)". The most important defects of the SVM for the project cost (*Y2*) and construction progress (*Y3*) classification models were "failure to implement a quality control checklist (Q76)" and "debris on concrete surface (W5)", respectively.

ML had a wide variety of classifiers, and each classifier had its own advantages and disadvantages. The prediction results of the classifier were related to the characteristics of the data to be classified, for example, the size of a dataset, the type of category, and dimensions. There was no single classifier that could have a perfect classification effect for all given problems. Thus, it was necessary to further analyze and compare the performance of the classifier based on various data training and testing results to determine the appropriate classification model. This research developed an optimized model for classifying project features and offered a comprehensive comparison among the effectiveness of seven ML algorithms. Therefore, project managers will be able to comprehend classification models and defects and use this to identify the most appropriate algorithm for classifying various project features.

Due to the parameter settings, operating principles and certain properties can limit the application of classification models. Furthermore, the feature conditions of tested data, for example, sample characteristics and construction types, may differ. The selection of appropriate models for project features classification may be discussed in the future. The imbalance of category samples is not the main source of classification difficulties. The reason behind it requires a more detailed observation of the data distribution, and the behavior of the model during the training process. Moreover, future studies must consider the goal of the research and data characteristics to determine the appropriate model to use.

Funding: The study was generously supported by the Ministry of Science and Technology, Taiwan (Contract No. MOST 109-2222-E-145-001). The authors are grateful to them for their financial support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest that might be perceived as affecting the objectivity of this study.

References

1. Meng, X. The effect of relationship management on project performance in construction. *Int. J. Proj. Manag.* **2012**, *30*, 188–198. [CrossRef]
2. Karim, K.; Marosszeky, M.; Davis, S. Managing subcontractor supply chain for quality in construction. *Eng. Constr. Archit. Manag.* **2006**, *13*, 27–42. [CrossRef]
3. Mills, A.; Love, P.E.D.; Williams, P. Defect costs in residential construction. *J. Constr. Eng. Manag.* **2009**, *135*, 12–16. [CrossRef]
4. Georgiou, J. Verification of a building defect classification system for housing. *Struct. Surv.* **2010**, *28*, 370–383. [CrossRef]
5. Ahzahar, N.; Karim, N.A.; Hassan, S.H.; Eman, J. A study of contribution factors to building failures and defects in construction industry. *Procedia Eng.* **2011**, *20*, 249–255. [CrossRef]
6. Forcada, N.; Macarulla, M.; Love, P.E.D. Assessment of residential defects at post-handover. *J. Constr. Eng. Manag.* **2013**, *139*, 372–378. [CrossRef]
7. Sinha, S.K.; Fieguth, P.W. Neuro-fuzzy network for the classification of buried pipe defects. *Autom. Constr.* **2006**, *15*, 73–83. [CrossRef]
8. Cheng, Y.; Yu, W.D.; Li, Q. A-based multi-level association rule mining approach for defect analysis in the construction industry. *Autom. Constr.* **2015**, *51*, 78–91. [CrossRef]
9. Elmasry, M.; Zayed, T.; Hawari, A. Defect based deterioration model for sewer pipelines using Bayesian belief networks. *Can. J. Civ. Eng.* **2017**, *44*, 675–690. [CrossRef]
10. Lin, C.L.; Fan, C.L. Examining association between construction inspection grades and critical defects using data mining and fuzzy logic. *J. Civ. Eng. Manag.* **2018**, *24*, 301–315. [CrossRef]
11. Okazaki, Y.; Okazaki, S.; Asamoto, S.; Chun, P.J. Applicability of machine learning to a crack model in concrete bridges. *Comput. Aided Civ. Inf.* **2020**, *35*, 775–792. [CrossRef]
12. Bu, G.; Lee, J.; Guan, H.; Loo, Y.; Blumenstein, M. Prediction of long-term bridge performance: Integrated deterioration approach with case studies. *J. Perform. Constr. Fac.* **2015**, *29*, 4014089. [CrossRef]
13. Lee, S.; Han, S.; Hyun, C. Analysis of causality between defect causes using association rule mining. *Int. J. Civ. Environ. Struct. Constr. Archit. Eng.* **2016**, *10*, 654–657.
14. Macarulla, M.; Forcada, N.; Casals, M.; Gangolells, M. Standardizing housing defects: Classification, validation, and benefits. *J. Constr. Eng. Manag.* **2013**, *139*, 968–976. [CrossRef]
15. Das, S.; Chew, M.Y.L. Generic method of grading building defects using FMECA to improve maintainability decisions. *J. Perform. Constr. Fac.* **2011**, *25*, 522–533. [CrossRef]
16. Rodrigues, J.; Folgado, D.; Belo, D.; Gamboa, H. SSTS: A syntactic tool for pattern search on time series. *Inform. Process. Manag.* **2019**, *56*, 61–76. [CrossRef]
17. Barbosa, M.W.; Vicente, A.C.; Ladeira, M.B.; Oliveira, M.P.V. Managing supply chain resources with big data analytics: A systematic review. *Int. J. Logist. Res. Appl.* **2018**, *21*, 177–200. [CrossRef]
18. Altincay, H.; Ergun, C. Clustering based undersampling for improving speaker verification decisions using AdaBoost. *Lect. Notes Comput. Sci.* **2004**, *3138*, 698–706.
19. Shirazi, F.; Mohammadi, M. A big data analytics model for customer churn prediction in the retiree segment. *Int. J. Inform. Manag.* **2019**, *48*, 238–253. [CrossRef]
20. Chien, C.; Wang, W.; Cheng, J. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Syst. Appl.* **2017**, *33*, 192–198. [CrossRef]

21. Tayefia, M.; Tajfard, M.; Saffar, S.; Hanachi, P.; Amirabadizadeh, A.R.; Esmaeily, H.; Taghipour, A.; Ferns, G.A.; Moohebati, M.; Ghayour-Mobarhan, M. Hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. *Comput. Methods Prog. Biomed.* **2017**, *141*, 105–109. [CrossRef] [PubMed]
22. Mistikoglu, G.; Gerek, I.H.; Erdis, E.; Usmen, P.E.M.; Cakan, H.; Kazan, E.E. Decision tree analysis of construction fall accidents involving roofers. *Expert Syst. Appl.* **2015**, *42*, 2256–2263. [CrossRef]
23. Arditi, D.; Pulket, T. Predicting the outcome of construction litigation using boosted decision trees. *J. Comput. Civ. Eng.* **2005**, *19*, 387–393. [CrossRef]
24. Shin, Y.; Kim, T.; Cho, H.; Kang, K.I. A formwork method selection model based on boosted decision trees in tall building construction. *Autom. Constr.* **2012**, *23*, 47–54. [CrossRef]
25. Murphy, K. The Bayes net toolbox for Matlab. *Comp. Sci. Stat.* **2001**, *33*, 1024–1034.
26. Straub, D.; Kiureghian, A.D. Bayesian network enhanced with structural reliability methods: Methodology. *J. Constr. Eng. Manag.* **2010**, *136*, 1248–1258. [CrossRef]
27. Ma, Y.F.; Wang, L.; Zhang, J.R.; Xiang, Y.B.; Liu, Y.M. Bridge remaining strength prediction integrated with Bayesian network and in situ load testing. *J. Bridge Eng.* **2014**, *19*, 4014037. [CrossRef]
28. Tesfamariam, S.; Martín-Pérez, B. Bayesian belief network to assess carbonation-induced corrosion in reinforced concrete. *J. Constr. Eng. Manag.* **2008**, *20*, 707–717. [CrossRef]
29. Maeda, K.; Takahashi, S.; Ogawa, T.; Haseyama, M. Distress classification of road structures via adaptive Bayesian network model selection. *J. Comput. Civ. Eng.* **2017**, *31*, 4017044. [CrossRef]
30. Tam, C.M.; Tong, T.K.L.; Lau, T.C.T.; Chan, K.K. Diagnosis of prestressed concrete pile defects using probabilistic neural networks. *Eng. Struct.* **2004**, *26*, 1155–1162. [CrossRef]
31. Petroutsatou, K.; Georgopoulos, E.; Lambropoulos, S.; Pantouvakis, J.P. Early cost estimating of road tunnel construction using neural networks. *J. Constr. Eng. Manag.* **2012**, *138*, 679–687. [CrossRef]
32. Marzouk, M.; Amin, A. Predicting construction materials prices using fuzzy logic and neural networks. *J. Constr. Eng. Manag.* **2013**, *139*, 1190–1198. [CrossRef]
33. Jafarzadeh, R.; Ingham, J.M.; Wilkinson, S.; Gonzalez, V.; Aghakouchak, A.A. Application of artificial neural network methodology for predicting seismic retrofit construction costs. *J. Constr. Eng. Manag.* **2014**, *140*, 4013044. [CrossRef]
34. Ayhan, B.U.; Tokdemir, O.B. Accident analysis for construction safety using latent class clustering and artificial neural networks. *J. Constr. Eng. Manag.* **2020**, *146*, 4019114. [CrossRef]
35. Pereira, E.; Ali, M.; Wu, L.; Abourizk, S. Distributed simulation–based analytics approach for enhancing safety management systems in industrial construction. *J. Constr. Eng. Manag.* **2020**, *146*, 4019091. [CrossRef]
36. Bai, L.; Wang, Z.; Wang, H.; Huang, N.; Shi, H. Prediction of multiproject resource conflict risk via an artificial neural network. *Engineering, Eng. Constr. Archit. Manag.* **2021**, *28*, 2857–2883. [CrossRef]
37. Park, J.K.; Hossain, S.; Oh, J.; Yoo, H.; Kim, H. Assessment of risk potential due to underground box structure installation employing ANN model and field experimental approaches. *J. Perform. Constr. Fac.* **2020**, *34*, 4020057. [CrossRef]
38. Murugan, S.B.; Sundar, M.L. Investigate safety and quality performance at construction site using artificial neural network. *J. Intell. Fuzzy Syst.* **2017**, *33*, 2211–2222. [CrossRef]
39. Cheng, M.Y.; Roy, A.F.V. Evolutionary fuzzy decision model for construction management using support vector machine. *Expert Syst. Appl.* **2010**, *37*, 6061–6069. [CrossRef]
40. Li, G.; Zhao, X.; Du, K.; Ru, F.; Zhang, Y. Recognition and evaluation of bridge cracks with modified active contour model and greedy search-based support vector machine. *Autom. Constr.* **2017**, *78*, 51–61. [CrossRef]
41. Hadjidemetriou, G.M.; Vela, P.A.; Christodoulou, S.E. Automated pavement patch detection and quantification using support vector machines. *J. Comput. Civ. Eng.* **2018**, *32*, 4017073. [CrossRef]
42. Liu, C.; Liu, C.; Liu, C.; Huang, X.; Miao, J.; Xu, W. Fire damage identification in RC beams based on support vector machines considering vibration test. *KSCE J. Civ. Eng.* **2019**, *23*, 4407–4416. [CrossRef]
43. Chen, P.H.; Shen, H.K.; Lei, C.Y.; Chang, L.M. Support vector machine based method for automated steel bridge rust assessment. *Autom. Constr.* **2012**, *23*, 9–19. [CrossRef]
44. Taffese, W.Z.; Sistonen, E. Machine learning for durability and service-life assessment of reinforced concrete structures: Recent advances and future directions. *Autom. Constr.* **2017**, *77*, 1–14. [CrossRef]
45. Chae, M.J.; Abraham, D.M. Neuro-fuzzy approaches for sanitary sewer pipeline condition assessment. *J. Comput. Civ. Eng.* **2001**, *15*, 4–14. [CrossRef]
46. Cheng, Y.M.; Leu, S.S. Integrating data mining with KJ method to classify bridge construction defects. *Expert Syst. Appl.* **2011**, *38*, 7143–7150. [CrossRef]
47. Gui, G.; Pan, H.; Lin, Z.; Li, Y.; Yuan, Z. Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection. *KSCE J. Civ. Eng.* **2017**, *21*, 523–534. [CrossRef]
48. Lin, C.L.; Fan, C.L. Evaluation of CART, CHAID, and QUEST algorithms: A case study of construction defects in Taiwan. *J. Asian Archit. Build.* **2019**, *18*, 539–553. [CrossRef]
49. Fan, C.L. Defect risk assessment using a hybrid machine learning method. *J. Constr. Eng. Manag.* **2020**, *146*, 4020102. [CrossRef]
50. Lee, M.C. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Syst. Appl.* **2009**, *36*, 10896–10904. [CrossRef]

51. Chou, J.S.; Tsai, C.F.; Lu, Y.H. Project dispute prediction by hybrid machine learning techniques. *J. Civ. Eng. Manag.* **2013**, *19*, 505–517. [CrossRef]
52. Gondia, A.; Siam, A.; El-Dakhakhni, W.; Nassar, A.H. Machine learning algorithms for construction projects delay risk prediction. *J. Constr. Eng. Manag.* **2020**, *146*, 4019085. [CrossRef]
53. Kifokeris, D.; Xenidis, Y. Risk source-based constructability appraisal using supervised machine learning. *Autom. Constr.* **2019**, *104*, 341–359. [CrossRef]
54. Chen, J.J.; Chen, E.E.; Zhao, W.; Zou, W. Statistics in big data. *J. Chin. Stat. Assoc.* **2015**, *53*, 186–202.
55. Piramuthu, S. Input data for decision trees. *Expert Syst. Appl.* **2008**, *34*, 1220–1226. [CrossRef]
56. Han, J.; Kamber, M. *Data Mining Concept and Technology*; Morgan Kaufmann: San Francisco, CA, USA, 2001.
57. Breiman, L.; Friedman, J.H.; Olshen, R.J.; Stone, C.J. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984.
58. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *Appl. Stat.* **1980**, *29*, 119–127. [CrossRef]
59. Loh, W.; Shih, Y. Split selection methods for classification trees. *Stat. Sin.* **1997**, *7*, 815–840.
60. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
61. Rafiei, M.H.; Adeli, H. Novel machine-learning model for estimating construction costs considering economic variables and indexes. *J. Constr. Eng. Manag.* **2018**, *144*, 4018106. [CrossRef]
62. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
63. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, NY, USA, 1995.
64. Halfawy, M.R.; Hengmeechai, J. Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine. *Autom. Constr.* **2014**, *38*, 1–13. [CrossRef]
65. Zhu, Z.; Brilakis, I. Parameter optimization for automated concrete detection in image data. *Autom. Constr.* **2010**, *19*, 944–953. [CrossRef]
66. Arlot, S. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
67. Neville, P.G. *Decision Trees for Predictive Modeling*; SAS Institute Inc.: Cary, NC, USA, 1999.
68. Yan, Y.; Liu, R.; Ding, Z.; Du, X.; Chen, J.; Zhang, Y. A Parameter-free cleaning method for smote in imbalanced classification. *IEEE Access* **2019**, *7*, 23537–23548. [CrossRef]
69. Salehi, H.; Burgueño, R. Emerging artificial intelligence methods in structural engineering. *Eng. Struct.* **2018**, *171*, 170–189. [CrossRef]
70. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
71. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2012.