Development of physical property prediction models for polypropylene composites with optimizing random forest hyperparameters

Chonghyo Joo1,2 | Hyundo Park1,3 | Jongkoo Lim4 | Hyungtae Cho1 | Junghwan Kim1

1Green Materials and Processes R&D Group, Korea Institute of Industrial Technology, Jung-gu, Ulsan, Republic of Korea
2Department of Chemical Engineering, Konkuk University, Gwangjin-gu, Seoul, Republic of Korea
3Department of Chemical and Biomolecular Engineering, Yonsei University, Seodaemun-gu, Seoul, Republic of Korea
4Research & Development Center, GS Caltex Corporation, Yuseon-gu, Daejeon, Republic of Korea

Correspondence
Junghwan Kim, Green Materials and Processes R&D Group, Korea Institute of Industrial Technology, 55, Jongga-ro, Jung-gu, Ulsan 44413, Republic of Korea.
Email: kjh31@kitech.re.kr

Abstract
The physical properties required in polypropylene composites (PPCs) vary depending on the purpose of use. In the manufacturing of PPCs, it is crucial to determine the types and quantities of numerous reinforcements to meet the required physical properties. Owing to industrial complexity, most PPC manufacturers produce the composites repeatedly until the desired physical properties are obtained. Hence, to reduce trial and error, we developed prediction models for the physical properties of PPCs based on commercial recipe data. The recipe data included information about five physical properties of composites manufactured using 90 materials. In complex industrial environments, because one recipe is usually composed of 2–12 materials, numerous combinations of data sets are created. It causes the lack of the same material combination data sets and thus makes it difficult to develop a good

Abbreviations: ANN, artificial neural network; CN, child node; DT, decision tree; FM, flexural modulus; FS, flexural strength; HPO, hyperparameter optimization; MAE, mean absolute error; MAPE, mean absolute percentage error; MDI, mean decrease in impurity; MI, melting index; ML, machine learning; NRMSE, normalized root mean-squared error; PN, parent node; PP, polypropylene; $R^2$, coefficient of determination; RF, random forest; RMSE, root mean-squared error; RSS, residual sum of squares; SG, specific gravity; TS, tensile strength; VI, variable importance.

Chonghyo Joo and Hyundo Park contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. International Journal of Intelligent Systems Published by Wiley Periodicals LLC

Int J Intell Syst. 2022;37:3625–3653.
performance model. Therefore, a novel categorization process is suggested as data preprocessing to overcome the data imbalance problem. The models for predicting the five physical properties (flexural strength, melting index, tensile strength, specific gravity, and flexural modulus) were developed using random forest, and the performance of the prediction models was improved via hyperparameter optimization. Furthermore, the effects of the materials on the performance of the models were numerically described through variable importance analysis. Finally, a software was developed to implement the prediction models in the industry. The software was applied to a commercial composite and achieved high accuracy, demonstrating the effectiveness of this study. Thus, the software suggests decision-making solutions to save cost and time by reducing the trial and error in the industrial environment with high complexity.

**KEYWORDS**
hyperparameter optimization, physical properties, polypropylene composites, prediction model, random forest

## 1 INTRODUCTION

A composite material is defined as a mixture (complex) of two or more materials. Generally, the mixture has a superior structure in performance and function than that of each component material. The general structure of the composite material is composed of a matrix, which serves as the bonding material, and a filler, which serves as reinforcement. One of the representative composite materials is the composite resin. Among the composite resin, polypropylene composites (PPCs) are widely used in various areas and in numerous applications, such as electrical equipment, laboratory equipment, textiles, reusable containers, aerospace structures, infrastructure, pipes, and tanks.1–7 PPCs have many merits with flowability, mechanical characteristics, weatherability, and chemical resistance, and are economical considering cost aspects.8

Because of the versatility of PPCs, the required physical properties are countless and vary depending on the purpose of end products. At present, most manufacturers in the composite resin industry attempt to meet the specifications of PPC products by trial and error. In other words, the manufacturing and testing of the composites are performed repeatedly until the desired properties are obtained. It is impossible to predict the physical properties of the PPC owing to the absence of theoretical or mathematical models about these physical properties; moreover, in complex industrial environments, because one recipe is usually composed of 2–12 materials, numerous combinations of data sets are created. They are therefore forced to utilize a similar recipe with the desired specifications first to produce the PPCs; thereafter, some raw
materials are added repeatedly until the required physical properties are obtained. This trial-and-error process reduces the productivity of the industry.

Materials 4.0, introduced by Jose and Ramakrishna, refers to the fourth paradigm of materials research, such as prediction of properties, design, and discovery of novel materials. The authors insisted that Materials 4.0 could use data to overcome human limitations, such as trial and error, and one of the methods of Materials 4.0 is machine learning (ML). Several materials studies using ML have been conducted successfully. Kazi et al. developed an artificial neural network (ANN) model for fiber-reinforced polymeric composites to satisfy required target properties while achieving the target filler content of composites. The ANN model also played a major role in determining the optimal filler content of cotton fabrics. Jiang et al. devised a melt index prediction model using a relevance vector machine (RVM) based on data from an actual polypropylene (PP) production plant for a reliable estimation of the melting index (MI) to control the quality of PP. The RVM-based model achieved a high MI prediction accuracy using nine process variables, such as process temperature, pressure, and flow rates. Ouyang et al. reviewed the applications of ML in studying and optimizing the thermal properties of materials. The authors found that ML plays an important role in the study of thermal transport in complex materials and can provide valuable insights into the material information without the significant computational costs of traditional simulation.

Recently, ML-based prediction models are being developed in various industries as well as in the materials industry. Coley et al. developed a prediction model using a convolutional neural network for four properties of molecules: aqueous solubility, octanol solubility, melting point, and toxicity. They demonstrated the use of neural network-based models that do not rely on exhaustive molecular descriptor calculations or experimental parameters. Kwon et al. used an ML-based prediction model to predict the temperature of the distillation column for controlling a chemical process. They optimized the operating conditions of the distillation process using the prediction model reduced steam consumption by approximately 14% using the prediction model. Li et al. used the Legendre moments algorithm and rotation forest to predict drug–target interactions. Drug–target interactions can be identified by traditional experimental methods; however, they are time-consuming, tedious, and expensive. To solve this problem, they proposed a prediction model with a drug molecular structure and protein sequence, and the proposed model proved superior to comparative methods. Tao et al. attempted to accelerate the discovery of efficient perovskite photocatalysts for photocatalytic water splitting using ML. The authors used the Pearson correlation coefficient to compare the performances of different algorithms, including gradient boosting, support vector machine, ANN, and random forest (RF). They employed a structural-property model with the highest accuracy for the rate of hydrogen production and the proper bandgap; thereafter, they developed user-friendly web servers to predict them, which could provide guidance for material design and optimization. Wang et al. developed a prediction model for the concentrations and spatial distributions of six heavy metals (Pb, Cd, Cr, As, Hg, and Zn) in soil using RF and verified that the developed model had a better performance than that of the existing land use regression model.

However, ML-based prediction models have an overfitting problem in limited industrial environments with a high complexity, such as the composite resin industry. RF is an ensemble method for learning multiple decision trees (DTs) and is used in various problems, such as classification and regression analysis. RF is also an effective algorithm for overcoming the problem of model overfitting in industry under limited resources. The superior predictive performance of RF has been verified through comparisons with many algorithms. Nguyen et al. used ML for the implementation of catalyst informatics. They developed an ML-prediction model for the C₂ yield of catalysts using atomic number for the catalyst design under
experimental conditions, such as temperature, pressure, and flow rates; here, RF achieved the best predictive performance among the algorithms (linear regression, least-squares linear regression, support vector regression with linear kernel, kernel ridge, and support vector regression) used.22 Watt et al. developed a dothistroma needle blight prediction model using environmental variables, such as relative humidity and compared the performance of the developed prediction models. The comparison showed that prediction models based on RF have the best prediction performance among multiple regression, XGBoost, and RF.23 Zekić-Sušac et al. predicted the energy cost of a public building using ANN, and RF to create an efficient model that can predict the cost of energy consumption and extract important features; the prediction performance of the RF algorithm was superior.24

Previous studies have identified problem-solving by developing ML-based prediction models. Moreover, RF algorithms performed better than other algorithms in the utilization of industrial data. Table 1 presents a summary of the proposed prediction models in the composite field literature. Despite the many contributions to developing prediction models for composites, several challenges remain to be addressed. The industry mostly has a complex environment; for example, it is difficult to develop a robust prediction model for PPCs data because the data are too biased. PPCs manufacturing data are organized using recipe data consisting of numerous materials, combinations, and compositions. However, most studies have focused on predicting properties using a few materials composed of only one combination rather than a recipe base, which is a mixture of various kinds of raw materials. Occasionally, these studies achieve high predictive performance on some data sets except data sets from complex industries. Moreover, some studies only focused on the prediction performance of the model and ignored the overfitting problem.

To solve these problems, we developed an RF-based model to predict the physical properties of PPCs to be manufactured based on numerous materials, material combinations, and compositions. A novel preprocessing categorization procedure is proposed to solve the overfitting problem. The prediction model is based on real commercial recipe data, and it uses the type and composition of the materials to predict the physical property of the PPC. We also developed a software that uses the prediction models for industrial applications, and it suggests decision-support solutions for industrial applications to save cost and time by reducing trial and error.

The novelty and major contributions of this study are as follows:

- This study proposed a framework for developing the prediction model by training recipe-based data in limited industrial environments with a high complexity.
- A novel categorization process is suggested as data preprocessing to overcome the overfitting issue due to the data imbalance problem.
- This is the first study that predicts the physical properties of PPCs with the numerous types and compositions of materials simultaneously.
- A user-friendly software is developed for implementing the prediction models in industry. The software has the model upgrade system by updating the new recipe data from the manufacturer to the industry; thus, the model performance is continuously intensified.

Figure 1 shows an overview of the development process of the physical property prediction model for PPCs. The prediction model using RF was developed by PPC recipe data, data preprocessing, and hyperparameter optimization (HPO). The developed prediction model was evaluated through evaluation criteria, and the material-specific impact was analyzed through variable importance (VI) analysis. Finally, the software containing five physical property prediction models was developed.
| References | ML methods | Industrial fields | Number of data/categories | Descriptions |
|------------|------------|-------------------|---------------------------|--------------|
| Kazi et al. \(^{10}\) | ANN        | Cotton fiber PP composite | Nondisclosure | • This study predicts the amount of targeted filler content with target properties  
  • This model predicts the weight percentage of only one material |
| Jiang et al. \(^{11}\) | MPSO-RVM   | PP composite      | 85/1                     | • This study focuses on the relation between the process operating conditions and PP property  
  • This study proposes only how to control the operational conditions for the property |
| Huang et al. \(^{10}\) | ANN SVM    | Cement composite  | 114/1                    | • This study predicts the properties of cement composites and explores the relationships among them  
  • This study does not include more influential factors in the model because of the limitations of the data |
| Lu et al. \(^{31}\) | MOGWO-SVM WOA-SVM PSO-SVM SVM | Building material | 207/1                    | • This model considers the prediction accuracy and stability of the model simultaneously  
  • This model has a problem of being less likely to be utilized because of the limitations and lack of the data set |
| Daghigh et al. \(^{32}\) | KNNR       | Bio-nano-composite | 50/16                    | • This study predicts the properties of composites using the compositions  
  • This model has a problem of being less likely to be utilized because of the limitations and lack of the data set |
| Qi et al. \(^{33}\) | PSO-ANN    | Cement composite  | 396/1                    | • This study predicts CPB strength using the PSO-ANN method  
  • This study has a limitation of omitting influential variables |

Abbreviations: ANN, artificial neural network; CPB, cemented paste backfill; KNNR, K-nearest neighbor regressor; ML, machine learning; MOGWO, multiobjective grey wolf optimizer; MPSO, modified particle swarm optimization; PP, polypropylene; PSO, particle swarm optimization; RVM, relevance vector machine; SVM, support vector machine; WOA, whale optimization algorithm.
The rest of this paper is organized as follows. Section 2 describes the RF algorithm, model performance evaluation criteria, and gridsearch. Section 3 details the process of developing the physical property prediction models. Section 4 presents the predictive performance of the models and the VI analysis results. Section 5 describes the software developed to incorporate the five proposed physical property prediction models. Finally, Section 6 summarizes the expected effectiveness of and utilization plan for the developed models and VI analysis results.

2 | PRELIMINARIES

2.1 | Random forest

RF is an ensemble learning method used for classification and regression analysis, first introduced in 2001 by Breiman. RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Relevant to the concept of RF is the DT, which is an ML method that connects the observed value for an item with the target value and has been used effectively in supervised classification learning. Because overfitting may occur when only one DT is used, RF which is composed of multiple DTs has been applied in recent research. Using the average of prediction results from multiple DTs or multiple prediction results in this

FIGURE 1 Overview of a prediction model using random forest. FS, flexural strength; PP, polypropylene [Color figure can be viewed at wileyonlinelibrary.com]
manner is called the ensemble method, which can solve overfitting problems because the generalization performance is enhanced compared with that of using only one DT.35

The process of forming a DT is as follows. First, \( n \) training data are randomly selected in the training data set. This process is called "bootstrap" and improves the stability and accuracy of the algorithm. Afterward, \( d \) feature values to be used as input data for the regression are selected from the selected \( n \) data (for the PPC data, the wt\% of the used material corresponds to the feature value). Finally, a DT is formed and trained using the \( n \) training data and \( d \) input data.

Data classification criteria are required to form a DT. The better the criteria, the higher the performance of the DT. This is because better data classification criteria lower data errors in a node. To choose the best data classification criteria, the DT calculates and compares the residual sum of squares (RSS) of all the feature values for every node. The criteria with the smallest RSS are selected for each node; subsequently, the nodes are divided by the criteria, and a DT is formed. When this process is repeated \( k \) times, an RF with \( k \) DTs is created. RSS and \( \hat{y} \) are calculated as shown in Equations (1) and (2).36,37

\[
J = \frac{m_{\text{left}}}{m} \sum (y - \hat{y}_{\text{left}})^2 + \frac{m_{\text{right}}}{m} \sum (y - \hat{y}_{\text{right}})^2, \tag{1}
\]

\[
\hat{y} = \frac{1}{m} \sum y, \tag{2}
\]

where \( J \) is the RSS, \( m_{\text{left}} \) is the amount of data in the left node, \( m_{\text{right}} \) is the amount of data in the right node, and \( m \) is the amount of data in the node before classification. \( y \) is the output of data (in the case of PPC data, the physical property of the used material corresponds to \( y \)), and \( \hat{y} \) is the mean of \( y \) in each node.

RF has a number of hyperparameters, but we performed optimization only for three of these: (i) the number of DTs forming the RF; (ii) the maximum depth of the DT, which is a hyperparameter of the DT; and (iii) the minimum number of samples in a node. The number of DTs literally denotes the number of DTs forming an RF. The DTs in an RF have different criteria for segmenting data and the different characteristics of the DTs, such as the number of nodes. Because RF derives its final result by summarizing the results of multiple DTs, the larger the number of DTs, the higher the performance of the RF-based prediction model. However, if the number of DTs is excessively large, the prediction will consume much time, and the accuracy will drop. Hence, it is crucial to determine an appropriate number of DTs for high prediction performance.

2.2 Evaluation criteria

There are many evaluation criteria for the performance of an ML prediction model, including root mean-squared error (RMSE), coefficient of determination \((R^2)\), mean absolute error (MAE), and mean absolute percentage error (MAPE).38–40 We evaluated the performance of the RF-based prediction models using RMSE, normalized RMSE (NRMSE), and \( R^2 \). These three evaluation criteria are calculated as shown in Equations (3)–(5):41–43

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N}(S_{ia} - S_{ip})^2}{N}}, \tag{3}
\]
\[
NRMSE = \frac{RMSE}{\max(S_{ia}) - \min(S_{ia})},
\]
where \(N\) is the number of data points used for calculation; \(S_{ia}\) is the \(i\)th real value; \(S_{ip}\) is the \(i\)th predicted value; and \(\max(S_{ia})\) and \(\min(S_{ia})\) are the maximum and minimum of the actual values, respectively.

\[
R^2 = 1 - \frac{\sum_{i=1}^{N} (S_{ia} - S_{ip})^2}{\sum_{i=1}^{N} (S_{ia} - \bar{S}_{ia})^2},
\]

3 | PREDICTION MODEL DEVELOPMENT

In this section, we describe the process of developing the prediction models for five physical properties. Figure 2 shows a flowchart of the model development, results, and application. The flowchart is in the same order as the configuration of the study.

3.1 | Data set

Table 2 shows the PPC recipe data (from GS Caltex Corporation) used in this study. The total number of recipes was 852. The recipes included data for real PP products of GS Caltex Corporation; thus, the data were encoded with alphanumeric characters based on a nondisclosure agreement. A part of the encoded data is released as support information (100 recipes). A total of 90 types of materials, including 41 PPs (P001–P041), 18 fillers (F001–F018), 22 rubbers (R001–R022), and 9 other additives (OTH1–OTH9), were included. Each recipe represents the types and quantities of materials used to produce PPCs and the physical properties of the composites produced using these recipes. This information about recipes (types and quantities of materials) are the inputs for prediction models, and the properties of recipes are the outputs. Although not all coded materials

![Flowchart of the prediction model development. RF, random forest; VI, variable importance](wileyonlinelibrary.com)
can be disclosed, the PPs included homo-PP, impact PP, and high crystallinity PP; the filler included mica and CaCO₃; the rubber included ethylene-butane rubber and ethylene-octene rubber. Other materials included carbon black, retardant, and so forth.

The data set encompassed five physical properties of PPCs, which were tested with American Society for Testing and Materials (ASTM) methods. The five physical properties include flexural strength (FS), MI, tensile strength (TS), specific gravity (SG), and flexural modulus (FM). FS refers to the maximum force that can bend a material without distorting or fracturing it. The three-point bending method is generally used to measure FS. The test apparatus has a high maintenance cost and space limitations for installation and use, and it requires a high cost for producing the necessary specimen. The proposed model is therefore expected to help reduce this testing cost.

| Recipe number | PP    | Filler | Rubber | Others | Physical property |
|---------------|-------|--------|--------|--------|-------------------|
|               |       |        |        |        | FS    | MI    | TS    | SG   | FM   |
| 1             | P030: | F016:  | R010:  | OTH1:  | 30.8  | 14.8  | 21.1  | 0.947| 2303 |
|               | 64.78 | 1.98   | 2.97   | 0.99   |       |       |       |      |      |
|               | P017: | 7.91   | R011:  | OTH3:  | 10     |       |       |      |      |
| 2             | P006: | F002:  | R003:  | OTH1:  | 2.1   | 7.4   | 0.996 | 62   |      |
|               | 23.48 | 14.68  | 51.86  | 1.96   |       |       |       |      |      |
|               |       |        |        | OTH4:  |       |       |       |      |      |
|               |       |        |        | 3.91   |       |       |       |      |      |
|               |       |        |        | OTH6:  | 0.20  |       |       |      |      |
|               |       |        |        | OTH8:  | 3.91  |       |       |      |      |
| 3             | P014: | F016:  | R010:  | OTH1:  | 31    | 26.36 | 21.4  | 0.943| 1541 |
|               | 29.52 | 7.62   | 1.90   | 2.86   |       |       |       |      |      |
|               | P030: | 50.48  | R017:  | OTH7:  | 1.90  |       |       |      |      |
| 4             | P014: | F009:  | R010:  | OTH1:  | 31.6  | 27    | 21.6  | 0.944| 1503 |
|               | 28.57 | 7.62   | 1.90   | 2.86   |       |       |       |      |      |
|               | P030: | 52.38  | R017:  | OTH7:  | 1.90  |       |       |      |      |
| 5             | P024: | F003:  | R002:  | OTH8:  | 1.9   | 9.43  | 1.5   | 0.93 | 46   |
|               | 17.65 | 4.90   | 24.51  | 1.96   |       |       |       |      |      |
|               |       |        | R003:  |       |       |       |       |      |      |
|               |       |        | 14.71  |       |       |       |       |      |      |
|               |       |        | R005:  | 29.41  |       |       |       |      |      |
|               |       |        | R007:  | 6.86   |       |       |       |      |      |
| 848           | P019: | F007:  | R012:  | OTH1:  | 36    | 26.9  | 23.3  | 0.997| 1995 |
|               | 42.16 | 13.73  | 5.88   | 1.96   |       |       |       |      |      |
|               | P022: |        | R013:  |       |       |       |       |      |      |
|               | 27.45 |        | 4.90   |       |       |       |       |      |      |
|               | P039: |        |        |       |       |       |       |      |      |
|               | 3.92  |        |        |       |       |       |       |      |      |
| 849           | P019: | F013:  | R010:  | OTH1:  | 27.8  | 19    | 1.026 | 1639 |
|               | 59.29 | 19.76  | 19.76  | 0.99   |       |       |       |      |      |
|               |       |        |        | OTH6:  | 0.20  |       |       |      |      |
| 850           | P029: | F009:  | R021:  | OTH1:  | 46.7  | 15.06 | 29    | 1.185| 4388 |
|               | 9.70  | 38.80  | 4.85   | 1.94   |       |       |       |      |      |
|               | P030: | 43.65  |        |       |       |       |       |      |      |
| 851           | P018: | F016:  | R010:  | OTH1:  | 24.2  | 30.5  | 17.3  | 0.994| 1826 |
|               | 56.93 | 11.39  | 2.85   | 2.85   |       |       |       |      |      |
|               | F017: | 5.69   | R014:  | 0.19   |       |       |       |      |      |
|               | 5.69  |        | 5.69   |       |       |       |       |      |      |
|               | R018: | 12.33  |        | 2.08   |       |       |       |      |      |
| 852           | P029: | F009:  | R013:  | OTH1:  | 33.2  | 7.3   | 22.5  | 1.04 | 2129 |
|               | 59.69 | 21.52  | 9.78   | 1.96   |       |       |       |      |      |
|               |       |        | R017:  | OTH3:  | 2.02  |       |       |      |      |
|               |       |        | 6.85   |       |       |       |       |      |      |

Abbreviations: FM, flexural modulus; FS, flexural strength; MI, melting index; PP, polypropylene; SG, specific gravity; TS, tensile strength.
MI refers to the weight of the resin that flows through a capillary for 10 min at a certain load and temperature. The MI measurement of PPCs is critical because it is directly related to the physical properties of the polymer products. However, MI measurement, like the aforementioned testing method for FS, also requires much time and cost.\textsuperscript{45,46} TS refers to the maximum stress when a material is fractured by a tensile load. TS is one of the most common items among mechanical property tests of plastics, specifically those relevant to tensioning. Meanwhile, SG refers to the weight of a resin per unit volume and is a fundamental physical property of polymers. SG must be measured for recipe development because it affects the physical properties and processing conditions of the processed products. Lastly, FM is the ratio of stress to strain within the elastic limit when a flexural load is applied to a polymer. In general, the larger the crystallinity and MI, the better the FM becomes.

3.2 Preprocessing

We performed two preprocessing steps before data training. The first step was the removal of missing values, and the second step was categorization. Missing values, which refer to data with empty values, must be removed in the preprocessing step because, otherwise, data would not be classified properly during the categorization step. The recipe data used in this study contained many missing values for the physical properties, and thus the missing values for each property had to be removed. After the missing values were removed, categorization was performed to solve the data imbalance problem. The data imbalance problem means that the data have a biased distribution. In complex industrial environments, because one recipe is generally composed of 2–12 materials, numerous combinations of data sets are created. If the data set is split randomly, the prediction models could be trained using data having only a few combinations of materials. This explains the overfitting issue due to the data imbalance problem during the data training process. To solve this problem, we grouped together recipes that use the same materials, referring to this process as categorization. Figure 3 shows the categorization process. First, when a certain material was used in a recipe (material's wt\% > 0), it was marked with a T; otherwise (material's wt\% = 0), it was marked with an F for data set binning. These labels were then changed to a code consisting of T's and F's, thus assigning a unique code to each recipe. If the codes were the same, the recipes were classified into the same type; otherwise, the recipes were classified into different types. Table 3 lists the number of recipes and categories remaining for each property after the missing values were removed.

After categorization, each of the groups included 1–12 recipes, indicating that the groups had from one recipe at the minimum to 12 recipes at the maximum, and the combinations had recipes featuring different compositions of the same materials. To train a model on all the recipe types, the prediction model must be trained on at least one recipe from each group. In addition to training the model on all recipe types, to avoid overfitting issues, the data were divided into training, validation, and test data sets. However, the random data split method typically used for regression does not require the model to be trained for all types of recipes. Hence, we defined the rule based on the number of recipes belonging to each group, as shown in Figure 4A. For example, if there were five recipes in one group, as shown in Figure 4B, three recipes could be classified into the training data set, one recipe to the validation data set, and one recipe to the test data set, respectively. As a result, three data sets were created for training, validation, and testing by the rule. Thus, the data imbalance problem could be solved because the model could be trained based on all types of recipes; moreover, the overfitting issue could be solved by using the data sets. Table 4 shows the number of...
the training, validation, and test data sets after the data were divided based on physical property. The division ratio was approximately 6:2:2.

3.3 Hyperparameters optimization

Hyperparameters are one of the most important factors in developing prediction models using ML, and they are also crucial for RF prediction models. A hyperparameter is a parameter
used to specify an algorithm for composing an ML model or minimizing the loss function, and thus the calculation time or prediction performance of the model depends on the hyperparameters. However, hyperparameters do not have set values because they cannot be inferred directly from data training and can vary based on the used data. Therefore, hyperparameters must be selected based on previous experience or trial and error. This process of determining the optimal set of hyperparameters is called “hyperparameter tuning” or “hyperparameter optimization (HPO).”

TABLE 3 Number of recipes and categories available per physical property

| Physical property | Number of recipes | Number of categories |
|-------------------|-------------------|----------------------|
| FS                | 844               | 484                  |
| MI                | 521               | 301                  |
| TS                | 842               | 483                  |
| SG                | 839               | 486                  |
| FM                | 847               | 485                  |

Abbreviations: FM, flexural modulus; FS, flexural strength; MI, melting index; SG, specific gravity; TS, tensile strength.

FIGURE 4 Data splitting process: (A) rules for data splitting and (B) example [Color figure can be viewed at wileyonlinelibrary.com]
In this study, we used a gridsearch, which is a method of comparing prediction performance for combinations of hyperparameters and is capable of comparing all combinations simultaneously. Gridsearch is often used for the optimization of hyperparameters because it can optimize the hyperparameters if the prediction model, type of hyperparameter, and data set are available. In general, gridsearch compares $R^2$ (score) using the $k$-fold cross-validation method, which evaluates a data set by dividing it into random $k$ parts. However, because this study used categorized data, we calculated the score using a validation data set generated through data split after categorization instead of using the $k$-fold method. Figure 5 shows the process of the gridsearch used in this study. The gridsearch process is as follows:

(i) Select hyperparameters for a case study of the RF (types and range of values).
(ii) Generate $n$ prediction models by applying $n$ combinations of hyperparameters generated by the selection in (i).
(iii) Train $n$ prediction models using the training data, and calculate the score by verifying the prediction model with the validation data.
(iv) Select the combination of hyperparameters with the highest score by comparing the scores calculated in (iii).

Table 5 shows the total cases of HPO. The number of DTs was set to 20, the maximum depth of DT was set to 10, and the minimum number of samples in a node was set to 1, 2, and 3. Thus, the total number of hyperparameter combinations compared using gridsearch was 600. The combinations determined using gridsearch were represented as “(number of DTs, maximum depth of DT, minimum number of samples in a node).”

Figure 6 shows a three-dimensional (3D) graph of the gridsearch results for FS indicated as scores according to the number of DTs, maximum depth of DT, and minimum number of samples in a node. The $x$-axis, $y$-axis, and $z$-axis represent the number of DTs, maximum depth of DT, and minimum number of samples in a node, respectively. The dots in the graph represent all tested combinations of hyperparameters, whereas the colors of the dots indicate the score according to the combination of hyperparameters. The score here signifies the $R^2$ calculated via a comparison of the actual FS with the FS predicted by the prediction model using the validation data. The closer the color is to yellow, the higher the score, and the closer the color is to purple, the lower the score. The trend of the hyperparameters can be determined based on the changes in the colors of the dots. In Figure 6, as the minimum number of samples in a node decreases from 3 to 1, the yellow dots increase in number. Thus, the smaller the
minimum number of samples in a node, the higher the performance of the RF-based prediction model. Regardless of the minimum number of samples in a node, the smaller the number of DTs and the maximum depth of DT, the lower the score. However, the combination with the highest score is (15, 55, 1), suggesting that a large number of DTs and a large maximum depth of DT do not always result in a high score. In Figure 6, the combinations of hyperparameters with the highest and lowest scores are (15, 55, 1) and (10, 10, 3), respectively, which are indicated by red squares. To predict the five physical properties, five RF-based physical property prediction models are required. To optimize the hyperparameters of the remaining MI, TS, SG, and FM prediction models, we found the optimized hyperparameter combination with the

![Diagram](image-url)
highest score using the same method described previously. Figures 7–10 show the gridsearch results of the prediction model for each physical property. The other physical properties (MI, TS, SG, and FM) proceed with HPO in the same way as that of FS. Using gridsearch, we compared the performance of the models for 600 hyperparameter combinations for each physical property. The results showed that the performance of the prediction model was changed by the combination of hyperparameters, and that the optimal hyperparameter combinations of the prediction model were different among the physical properties.
4 | RESULTS

4.1 | Model performance

After HPO, parity plots and evaluation metrics were used to determine the performance of each prediction model. Figure 11 shows parity plots for every property prediction results when the hyperparameter combinations are the worst and the best. In Figure 11, the x-axis represents the actual FS values of the test data set, whereas the y-axis represents the FS value predicted by...
the prediction model. The closer the blue dots to the straight line $y = x$ (red), the better the match between the actual and predicted values. When the distributions of dots are compared, the dots of the prediction model with the hyperparameter combination (15, 55, 1) are observed to be closer to the straight line $y = x$. In other words, the prediction model performance with the
combination (15, 55, 1) is higher than that with the combination (10, 10, 3). RMSE, NRMSE, and $R^2$ were then used to compare the performance of the prediction model based on objective values. The prediction model with the hyperparameter combination (10, 10, 3), which had the lowest score, had an RMSE of 10.3748, NRMSE of 0.0558, and $R^2$ of 0.8895. By contrast, the prediction model with the hyperparameter combination (15, 55, 1), which had the highest score, had an RMSE of 9.1163, NRMSE of 0.0491, and $R^2$ of 0.9147. Thus, the hyperparameter combination with the highest score had an RMSE and NRMSE lower by 1.2585 and 0.0067, respectively, and an $R^2$ higher by 0.0252 compared with those of the hyperparameter combination (10, 10, 3).
combination with the lowest score. Using the same method described for the FS parity plot, MI, TS, SG, and FM parity plots are shown in Figures 12–15, respectively.

Table 6 shows the results of evaluating the performance for each physical property based on the hyperparameter combinations with the lowest and highest scores, as obtained from the gridsearch results. As with the FS prediction model, the performance evaluations of these prediction models were performed using the test data set. The optimized hyperparameter combination for the MI prediction model was (10, 40, 1), of which the RMSE and NRMSE

*FIGURE 10  Gridsearch results for FM. DT, decision tree; FM, flexural modulus [Color figure can be viewed at wileyonlinelibrary.com]*
were lower by 0.9782 and 0.0173, respectively, and the $R^2$ value was higher by 0.0549 than those of the lowest-scoring combination (5, 35, 3). The optimized hyperparameter combination of the TS prediction model was (20, 55, 1), of which the RMSE and NRMSE values were lower by 0.6775 and 0.0047, respectively, and the $R^2$ value was higher by 0.0154 than those of the lowest-scoring combination (5, 10, 1). The optimized hyperparameter combination of the SG prediction model was (20, 55, 1), of which the RMSE and NRMSE values were lower by 0.0138 and 0.0276, respectively, and the $R^2$ value was higher by 0.0803 than those of the lowest-scoring combination (10, 10, 1). The optimized hyperparameter combination of the FM prediction model was (5, 55, 1), of which the RMSE and NRMSE values were lower by
103.2602 and 0.0073, respectively, and the $R^2$ value was higher by 0.0174 than those of the lowest-scoring combination (10, 10, 2).

### 4.2 VI analysis

The disadvantage of RF is that it is difficult to explain the relationship between input and output data because it is a black-box algorithm. To complement this shortcoming, VI, which is a value representing the input data that have an important role in prediction performance within the RF algorithm, was used. Therefore, the degree of effect of specific input data on
RMSE, NRMSE, and $R^2$ of the prediction models can be compared through VI analysis.\textsuperscript{51,52} Among the different types of VI, we used the mean decrease in impurity (MDI) importance, which is a representative VI. MDI importance represents the average of impurity reduction when input data are split individually, and it allows for a fast intuitive analysis.\textsuperscript{53} Figure 16 shows one of the DTs, which comprises the FS and MI prediction models. The first criteria of classification in the DTs are $X[44]$ and $X[21]$, which refer to F004 in Figure 16A and P022 in Figure 16B. In addition to F004 and P022, many types of materials could be used as the criteria for classification. Among them, the best criterion to classify data is referred to as a material.

**FIGURE 15** Parity plots for FM prediction: (A) hyperparameter combination (10, 10, 2), (B) hyperparameter combination (5, 55, 1). FM, flexural modulus [Color figure can be viewed at wileyonlinelibrary.com]

**TABLE 6** Evaluation of prediction models for tested physical properties

| Hyperparameters | Evaluation |
|-----------------|------------|
| | Number of DTs | Maximum depth of DT | Minimum number of samples in node | RMSE | NRMSE | $R^2$ |
| FS | 10 | 10 | 3 | 10.3748 | 0.0558 | 0.8895 |
| | 55 | 15 | 1 | 9.1163 | 0.0491 | 0.9147 |
| MI | 5 | 35 | 3 | 4.9429 | 0.0876 | 0.8461 |
| | 10 | 40 | 1 | 3.9647 | 0.0702 | 0.9010 |
| TS | 5 | 10 | 1 | 5.7666 | 0.0398 | 0.9304 |
| | 20 | 55 | 1 | 5.0891 | 0.0351 | 0.9458 |
| SG | 10 | 10 | 1 | 0.0320 | 0.0538 | 0.8816 |
| | 20 | 55 | 1 | 0.0182 | 0.0305 | 0.9619 |
| FM | 10 | 10 | 2 | 444.7644 | 0.0315 | 0.9577 |
| | 5 | 55 | 1 | 341.5042 | 0.0242 | 0.9751 |

Abbreviations: DT, decision tree; FM, flexural modulus; FS, flexural strength; MI, melting index; NRMSE, normalized root mean-squared error; RMSE, root mean-squared error; SG, specific gravity; TS, tensile strength.
with a high VI value. In this study, using VI analysis, we identified the materials with high VI values for every prediction model. Figure 17 shows the VI values for the top 10 materials that have the largest effects on the five tested physical properties of PPCs, that is, FS, MI, TS, SG, and FM. The VI value of the material that had the largest effect on the performance of the prediction model was set to 1, whereas the VI values of the other top nine materials were expressed as a value relative to 1. The VI value of the FS prediction model was the highest for F004, a type of filler, followed by the VI values of additives OTH5 and OTH1, at 0.01731 and
0.01621, respectively. By contrast, the VIs of materials other than F004 had very low values. Meanwhile, the VI value of the MI prediction model was the highest for P022, a type of PP, followed by P034 and R017, a type of rubber, at 0.60942 and 0.32166, respectively. Because six of the top 10 materials with high VI values were PP, PP is more important than filler, rubber, and other additives for the MI prediction models. The TS, SG, and FM prediction models have F004 as the material with the highest VI value similar to that of the FS prediction model. Consequently, the input and ratio of F004 are important to control the FS, TS, SG, and FM of the PPCs, and the selection of the type and ratio of PP are important to adjust the MI.

5 | APPLICATION: SOFTWARE FOR PHYSICAL PROPERTY PREDICTION

On the basis of our proposed prediction models for five physical properties, we developed a software for predicting the physical properties, using the PyInstaller library. The language of the software is python. The software simultaneously shows the five physical property prediction values of the PPC that would be generated from the composition value of the materials. Figure 18 shows the running screen of the software, which is composed of a field for entering the input composition (wt%) of each material, a field for the prediction results, and buttons, including “Predict” and “Cancel.”

![Graphical user interface of the software](wileyonlinelibrary.com)

**Figure 18** Graphical user interface of the software [Color figure can be viewed at wileyonlinelibrary.com]

**Table 7** Data for recipe 782

| Recipe number | PP      | Filler  | Rubber  | Others | Physical property |
|---------------|---------|---------|---------|--------|-------------------|
|               | FS      | MI      | TS      | SG     | FM               |
| 782           | P019: 59.29 | F013: 19.76 | R010: 19.76 | OTH1: 0.99 | 27.8 | 19 | 1.026 | 1639 |
|               | OTH6: 0.20 |         |         |        |                  |

Abbreviations: FM, flexural modulus; FS, flexural strength; MI, melting index; PP, polypropylene; SG, specific gravity; TS, tensile strength.
The software is simple to run. When the material composition (wt%) of the recipe is inputted, and the “Predict” button is clicked, the predicted physical properties of the resulting PPC are calculated and displayed on the screen. As an example, Table 7 shows the types, composition, and physical properties of the materials used in recipe 782, which was randomly selected from among the data not used for training. Recipe 782 is of a composite resin made of a combination of five materials: P019 (59.29 wt%), F013 (19.76 wt%), R010 (19.76 wt%), OTH1 (0.99 wt%), and OTH6 (0.2 wt%). Table 8 shows the predicted values and error rates for each physical property as calculated by the software for recipe 782. When the actual and predicted values were compared, the two values were highly similar, whereas the MI, a missing value, could also be predicted. Among the five physical properties, SG exhibited the highest accuracy, with the lowest error rate of 0.48%, whereas FS exhibited the highest error rate of 4.6%. In addition, the TS prediction model exhibited an error rate of 2.1% with the actual value, whereas the FM prediction model exhibited an error rate of 4.4%.

According to the results of the test, although slight errors exist for all predicted physical property values, these are expected to be solvable with an increase in the amount of training data. Furthermore, because the five physical property prediction models are independent of each other, they have an advantage in that if the prediction accuracy is low for a specific property, only the model of the corresponding property would need to be improved. When the developed software is applied to the actual composite-manufacturing industry, the physical properties of the composites can be predicted, thus saving the time and cost in the development of new recipes of composites with desired properties. Additionally, the software can be improved continuously by training the model with a newly developed recipe.

6 | DISCUSSIONS

Currently, most manufacturers in the industry meet the specifications of PPC products by a trial-and-error repetition of composite resin manufacturing and testing. This trial-and-error approach is adopted because of the lack of a theoretical or mathematical model for complex resin properties. We aimed to address this problem by developing a prediction model using ML. However, when existing research methods are used to develop prediction models for the composite resin industry data, it is difficult to apply them to industry because the combination and composition of numerous materials create infinitely many cases; moreover, the data are unevenly distributed and biased.

Therefore, in this study, we used the RF algorithm and a preprocessing method named categorization to prevent overfitting problems caused by data imbalance; we developed a model to predict the physical properties of PPCs to be produced by the type and composition of materials used in composite resins. The five physical property prediction models we developed achieved a predictive performance above $R^2$ of 0.9. The proposed models provide the physical properties of the materials used in composite resins.
property values of PPCs to be manufactured without the repetitive processes of “PPCs manufacturing” and “physical property testing.”

Additionally, we developed a physical property prediction software to facilitate the use of the proposed prediction models in industry. The software is a graphical user interface (GUI) that integrates the five prediction models to provide the property values of PPCs that will be manufactured; the inputs of the software are materials and composition values of PPCs manufacturing.

Our proposed prediction models and developed software enable us to know the physical property values before manufacturing PPCs, thereby reducing trial and error and saving time and costs incurred during PPCs manufacturing and testing. However, even if the physical property values of the PPCs to be manufactured using the software can be immediately known, the disadvantage involves repeatedly using the software until the desired physical property values are obtained. Therefore, we intend to conduct a study on a model that determines the recipe for producing PPCs with desired physical property values.

7 | CONCLUSION

In this study, we developed prediction models for five physical properties, namely, FS, MI, TS, SG, and FM of PPCs using actual recipe data from the industry. The performance of the prediction models was improved by optimizing the hyperparameters, which was performed via gridsearch for the number of DTs, maximum depth of DT, and minimum number of samples in a node, from several possible hyperparameters. The optimal combinations of hyperparameters calculated for these physical property prediction models were different from one another. Moreover, the performance indices of the prediction models, that is, RMSE, NRMSE, and $R^2$ values, were also different. Nonetheless, all five physical property prediction models exhibited high performance with an $R^2$ value of 0.9. Moreover, the degree of effect of each material used in the development of the prediction models was numerically expressed using VI analysis. The VI analysis result showed that the material with the highest VI was F004 for the FS, TS, SG, and FM prediction models, that is, all tested physical properties except MI. On the other hand, P022 exhibited the highest VI for the MI prediction model. Moreover, unlike with the other physical properties, six PPs, including P022, were included among the top 10 materials for the MI prediction model. A software that utilizes the prediction models was developed for industrial applications. The software application predicts and displays the physical property values of a PPC based on the types and compositions of its component materials. When the software is used in the actual composite-manufacturing industry, the manufacturing and testing process can reduce trial and error because the physical properties of the resulting composite can be predicted beforehand. Consequently, the software suggests decision-support solutions for industrial applications to save cost and time by reducing trial and error.

ACKNOWLEDGEMENTS

This study has been conducted with the support of the Korea Institute of Industrial Technology as “Development of AI Platform Technology for Smart Chemical Process (kitech JH-21-0005).”

AUTHOR CONTRIBUTIONS

Chonghyo Joo contributed to developing the model and the computational framework and analyzing the data, and writing the manuscript. Hyundo Park contributed to the design and implementation of the research, the analysis of the results, and the writing of the manuscript. Jongkoo Lim contributed to performing the experiment and devised the project. Hyungtae Cho
contributed to conceiving the study and was in charge of overall direction and planning, the analysis of the results, and the writing of the manuscript. Junghwan Kim contributed to conceiving the study and was in charge of overall direction and planning, discussing the results and commented on the manuscript, providing critical feedback, and helping shape the research, analysis, and manuscript.

ORCID

Chonghyo Joo [http://orcid.org/0000-0001-6486-8077]
Hyundo Park [http://orcid.org/0000-0001-5876-6867]
Jongkoo Lim [http://orcid.org/0000-0002-8654-1750]
Hyungtae Cho [http://orcid.org/0000-0002-7728-2793]
Junghwan Kim [http://orcid.org/0000-0002-2311-4567]

REFERENCES

1. Prokhorova IA, Bkkar M. Prediction of strength properties of elastic polymeric composite materials based on polypropylene woven tapes. Fibre Chem. 2020;52(3):173-178. https://doi.org/10.1007/s10692-020-10175-w
2. Ray K, Patra H, Swain AK, et al. Glass/jute/sisal fiber reinforced hybrid polypropylene polymer composites: fabrication and analysis of mechanical and water absorption properties. Mater Today Proc. 2020;33(Part 8):5273-5278. https://doi.org/10.1016/j.matpr.2020.02.964
3. Mahdi E, Dean A. The effect of filler content on the tensile behavior of polypropylene/cotton fiber and poly(vinyl chloride)/cotton fiber composites. Materials (Basel). 2020;13(3):753. https://doi.org/10.3390/ma13030753
4. Ismail H, Suryadiansyah. Thermoplastic elastomers based on polypropylene/natural rubber and polypropylene/recycle rubber blends. Polym Test. 2002;21(4):389-395. https://doi.org/10.1016/S0142-9418(01)00101-5
5. Ameli A, Nofer M, Wang S, Park CB. Lightweight polypropylene/stainless-steel fiber composite foams with low percolation for efficient electromagnetic interference shielding. ACS Appl Mater Interfaces. 2014;6(14):11091-11100. https://doi.org/10.1021/ami500445g
6. Zhang YP, Zhou CG, Sun WJ, et al. Injection molding of segregated carbon nanotube/polypropylene composite with enhanced electromagnetic interference shielding and mechanical performance. Compos Sci Technol. 2020;197(14):108253. https://doi.org/10.1016/j.compscitech.2020.108253
7. Dikobe DG, Luyt AS. Morphology and properties of polypropylene/ethylene vinyl acetate copolymer/wood powder blend composites. eXPRESS Polym Lett. 2009;3(3):190-199. https://doi.org/10.3144/expresspolymlett.2009.24
8. Nomura M, Shanmuga Ramanan SM, Arun S. Automobile bumpers. In: Comprehensive Composite Materials II. Elsevier; 2017:460-468. https://doi.org/10.1016/B978-0-12-803581-8.03962-X
9. Jose R, Ramakrishna S. Materials 4.0: materials big data enabled materials discovery. Appl Mater Today. 2018;10:127-132. https://doi.org/10.1016/j.apmt.2017.12.015
10. Kazi MK, Eljack F, Mahdi E. Optimal filler content for cotton fiber/PP composite based on mechanical properties using artificial neural network. Compos Struct. 2020;251:112654. https://doi.org/10.1016/j.composites.2020.112654
11. Jiang H, Xiao Y, Li J, Liu X. Prediction of the melt index based on the relevance vector machine with modified particle swarm optimization. Chem Eng Technol. 2012;35(5):819-826. https://doi.org/10.1002/ceat.201100437
12. Ouyang Y, Yu C, Yan G, Chen J. Machine learning approach for the prediction and optimization of thermal transport properties. Front Phys. 2021;16(4):43200. https://doi.org/10.1007/s11467-020-1041-x
13. Tao Q, Lu T, Sheng Y, Li L, Lu W, Li M. Machine learning aided design of perovskite oxide materials for photocatalytic water splitting. J Energy Chem. 2021;60:351-359. https://doi.org/10.1016/j.jechem.2021.01.035
14. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model. 2017;57(8):1757-1772. https://doi.org/10.1021/acs.jcim.6b00601
15. Kwon H, Oh KC, Choi Y, Chung YG, Kim J. Development and application of machine learning-based prediction model for distillation column. *Int J Intell Syst*. 2021;36(5):1970-1997. https://doi.org/10.1002/int.22368

16. Li Y, Liu X-z, You ZH, Li LP, Guo JX, Wang Z. A computational approach for predicting drug–target interactions from protein sequence and drug substructure fingerprint information. *Int J Intell Syst*. 2020;36:593-609. https://doi.org/10.1002/int.22332

17. Wang H, Yilihamu Q, Yuan M, Bai H, Xu H, Wu J. Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: a comparison of regression and random forest. *Ecol Indic*. 2020;119:106801. https://doi.org/10.1016/j.ecolind.2020.106801

18. Roh J, Jang S, Kim S, Cho H, Kim J. Steam trap maintenance-prioritizing model based on big data. *ACS Omega*. 2021;6(6):4408-4416. https://doi.org/10.1021/acsomega.0c05784

19. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324

20. Feng R, Grana D, Balling N. Imputation of missing well log data by random forest and its uncertainty analysis. *Comput Geosci*. 2021;152:104763. https://doi.org/10.1016/j.cageo.2021.104763

21. Talukdar S, Eibek KU, Akhter S, Ziaul S, Towfiqul Islam ARM, Mallick J. Modeling fragmentation probability of land-use and land-cover using the bagging, random forest and random subspace in the Teesta River Basin, Bangladesh. *Ecol Indic*. 2021;126:107612. https://doi.org/10.1016/j.ecolind.2021.107612

22. Nguyen TN, Nhat TTP, Takimoto K, et al. High-throughput experimentation and catalyst informatics for oxidative coupling of methane. *ACS Catal*. 2020;10(2):921-932. https://doi.org/10.1021/acscatal.9b04293

23. Watt MS, Tan AYS, Fraser S, Bulman LS. Use of advanced modelling methods to predict dothistroma needle blight on *Pinus radiata* at a fine resolution within New Zealand. *For Ecol Manage*. 2021;492:119226. https://doi.org/10.1016/j.foreco.2021.119226

24. Zeki-Sušac M, Has A, Knežević M. Predicting energy cost of public buildings by artificial neural networks, CART, and random forest. *Neurocomputing*. 2021;439:223-233. https://doi.org/10.1016/j.neucom.2020.01.124

25. Kapse S, Janwari S, Waghmare UV, Thapa R. Energy parameter and electronic descriptor for carbon based catalyst predicted using QM/ML. *Appl Catal B Environ*. 2021;286:119866. https://doi.org/10.1016/j.apcatb.2021.119866

26. Pegalajar MC, Ruiz LGB, Cuéllar MP, Rueda R. Analysis and enhanced prediction of the Spanish Electricity Network through big data and machine learning techniques. *Int J Approx Reason*. 2021;133:48-59. https://doi.org/10.1016/j.ijar.2021.03.002

27. Shukla K, Dadheech N, Kumar P, Khare M. Regression-based flexible models for photochemical air pollutants in the national capital territory of megalcity Delhi. *Chemosphere*. 2021;272:129611. https://doi.org/10.1016/j.chemosphere.2021.129611

28. Lei C, Deng J, Cao K, et al. A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob. *Fuel*. 2019;239:297-311. https://doi.org/10.1016/j.fuel.2018.11.006

29. Palmer DS, O’Boyle NM, Glen RC, Mitchell JBO. Random forest models to predict aqueous solubility. *J Chem Inf Model*. 2007;47(1):150-158. https://doi.org/10.1021/ci060164k

30. Huang JS, Liew JX, Liew KM. Data-driven machine learning approach for exploring and assessing mechanical properties of carbon nanotube-reinforced cement composites. *Compos Struct*. 2021;267:113917. https://doi.org/10.1016/j.compstruct.2021.113917

31. Lu H, Bebhahani S, Ma X, Iseley T. A multi-objective optimizer-based model for predicting composite material properties. *Constr Build Mater*. 2021;284:122746. https://doi.org/10.1016/J.CONBUILDMAT.2021.122746

32. Daghigh V, Lacy TE, Daghigh H, et al. Heat deflection temperatures of bio-nano-composites using experiments and machine learning predictions. *Mater Today Commun*. 2020;22:100789. https://doi.org/10.1016/J.MTCOMM.2019.100789

33. Qi C, Fourie A, Chen Q. Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill. *Constr Build Mater*. 2018;159:473-478. https://doi.org/10.1016/J.CONBUILDMAT.2017.11.006

34. Pavlov YL. *Random Forests*. Taylor & Francis Group. 2019:1-122. https://doi.org/10.1201/9780367816377-11

35. Johansson U, Boström H, Löfström T, Linusson H. Regression conformal prediction with random forests. *Mach Learn*. 2014;97(1-2):155-176. https://doi.org/10.1007/s10994-014-5453-0

36. Sega M, Xiao Y. Multivariate random forests. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2011;1(1):80-87. https://doi.org/10.1002/widm.12
37. 1.10. Decision Trees—scikit-learn 0.24.2 documentation. Accessed May 28, 2021. https://scikit-learn.org/stable/modules/tree.html#id2

38. Li H. Analysis on the construction of sports match prediction model using neural network. Soft Comput. 2020;24(11):8343-8353. https://doi.org/10.1007/s00500-020-04823-w

39. Rather AM, Agarwal A, Sastry VN. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst Appl. 2015;42(6):3234-3241. https://doi.org/10.1016/j.eswa.2014.12.003

40. Arthur E, Rehman HU, Fuller M, et al. Estimating Atterberg limits of soils from hygroscopic water content. Geoderma. 2021;381:114698. https://doi.org/10.1016/j.geoderma.2020.114698

41. Lang L, Liang T, Ai S, Tang X. An improved random forest algorithm and its application to wind pressure prediction. Int J Intell Syst. 2021;36(8):4016-4032. https://doi.org/10.1002/int.22448

42. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014;7(3):1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

43. Li CN, Shao YH, Zhao D, Guo YR, Hua XY. Feature selection for high-dimensional regression via sparse LSSVR based on Lp-norm. Int J Intell Syst. 2021;36(2):1108-1130. https://doi.org/10.1002/int.22334

44. Zhang X, Han X, Zhang Y, Liang J, Wang C, Liang J. Effect of holding time and pressure on the densification, microstructure and mechanical properties of hot pressed Al2O3-CA6-ZrO2/Ni multi-phase composites. J Alloys Compd. 2021;850:156783. https://doi.org/10.1016/j.jallcom.2020.156783

45. Zhang M, Liu X. A soft sensor based on adaptive fuzzy neural network and support vector regression for industrial melt index prediction. Chemom Intell Lab Syst. 2013;126:83-90. https://doi.org/10.1016/j.chemolab.2013.04.018

46. Bafna SS, Beall AM. A design of experiments study on the factors affecting variability in the melt index measurement. J Appl Polym Sci. 1997;65(2):277-288. https://doi.org/10.1002/(sici)1097-4628(19970711)65:2<277::aid-app8>3.0.co;2-l

47. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing. 2020;415:295-316. https://doi.org/10.1016/j.neucom.2020.07.061

48. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization: In: 24th International Conference on Neural Information Processing Systems. 2011;2546-2554. https://dl.acm.org/doi/10.5555/2986459.2986743

49. Xiao HB, Cao HT, Fan JH, Costantin D, Luo GY, Pei ZY. Efficient Fermi source identification with machine learning methods. Astron Comput. 2020;32:100387. https://doi.org/10.1016/j.ascom.2020.100387

50. Mezzatesta S, Torino C, De Meo P, Fiumara G, Vilasi A. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Comput Methods Programs Biomed. 2019;177:9-15. https://doi.org/10.1016/j.cmpb.2019.05.005

51. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. 2007;8(1):8-25. https://doi.org/10.1186/1471-2105-8-25

52. Zhang X, Wada T, Fujiwara K, Kano M. Regression and independence based variable importance measure. Comput Chem Eng. 2020;135:106757. https://doi.org/10.1016/j.compchemeng.2020.106757

53. Chen D, Chang N, Xiao J, Zhou Q, Wu W. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. Sci Total Environ. 2019;669:844-855. https://doi.org/10.1016/j.scitotenv.2019.03.151

SUPPORTING INFORMATION
Additional supporting information may be found in the online version of the article at the publisher’s website.