

Article

# An Advanced Soil Classification Method Employing the Random Forest Technique in Machine Learning

Chih-Yu Liu <sup>1,2</sup> , Cheng-Yu Ku <sup>1,2,\*</sup> , Ting-Yuan Wu <sup>1</sup> and Yun-Cheng Ku <sup>3</sup>

<sup>1</sup> Department of Harbor and River Engineering, National Taiwan Ocean University, Keelung 202301, Taiwan; 20452003@email.ntou.edu.tw (C.-Y.L.); 11252007@email.ntou.edu.tw (T.-Y.W.)

<sup>2</sup> Center of Excellence for Ocean Engineering, National Taiwan Ocean University, Keelung 202301, Taiwan

<sup>3</sup> Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taipei 106335, Taiwan; b11005105@mail.ntust.edu.tw

\* Correspondence: chkst26@mail.ntou.edu.tw

**Abstract:** Soil classification is essential for understanding soil properties and their suitability for conveying the characteristics of soil types. In this study, we present a prediction of soil classification using fewer soil variables by employing the random forest (RF) technique in machine learning. This study compiled the parameters outlined in the unified soil classification system (USCS), a widely used method for categorizing soils based on their properties and behavior. These parameters, encompassing grain size distribution, Atterberg limits, the coefficient of uniformity, and the coefficient of curvature, were defined within specific ranges to create a synthetic database for training the RF model. The importance of input factors in soil classification was assessed using the out-of-bag samples in RF. Through rigorous validation techniques, including cross-validation, the performance of the RF model is thoroughly assessed, demonstrating its capability to accurately evaluate soil classification. The findings indicate that the RF model presented in this study exhibits a promising alternative, providing automated and accurate classification based on soil data. Notably, the model indicates that the coefficients of uniformity and gradation are insignificant for soil classification and can predict soil types even when these factors are missing, a feat that traditional methods struggle to achieve.

**Keywords:** soil; unified soil classification system; random forest; grain size; Atterberg limits



**Citation:** Liu, C.-Y.; Ku, C.-Y.; Wu, T.-Y.; Ku, Y.-C. An Advanced Soil Classification Method Employing the Random Forest Technique in Machine Learning. *Appl. Sci.* **2024**, *14*, 7202. <https://doi.org/10.3390/app14167202>

Academic Editor: Marek Lefik

Received: 9 July 2024

Revised: 9 August 2024

Accepted: 13 August 2024

Published: 16 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Soil classification serves as a foundational component of geotechnical engineering, offering a structured system to organize and classify soils according to their physical and engineering attributes [1–3]. This systematic approach enables engineers and geologists to comprehend and communicate the properties and behaviors of various soil types, facilitating applications such as foundation design, construction planning, and environmental management [4–7]. Classical soil classification methods include the unified soil classification system (USCS), the United States department of agriculture (USDA) soil taxonomy, the American association of state highway and transportation officials (AASHTO) classification system, the Canadian system of soil classification, and the British soil classification system [8–14]. The USCS is a widely adopted method for classifying soils based on their physical and mechanical properties. It is extensively utilized in geotechnical engineering to determine soil behavior for construction projects, assess soil suitability, and design foundations. The AASHTO soil classification system, on the other hand, is mainly used to evaluate soils for highway construction and other transportation infrastructure. It aids in assessing soil suitability for use as subgrade material, predicting performance under load, and guiding construction practices. The AASHTO system focuses on soil behavior related to highway performance, including stability and drainage. Both the USCS and AASHTO system offer valuable frameworks for soil classification, each addressing specific

engineering needs. While these systems remain fundamental in soil classification, their effectiveness is significantly enhanced by modern computational techniques, which provide more accurate and efficient soil evaluations for various engineering applications.

The USCS stands out as one of the most extensively utilized soil classification frameworks globally. It groups soils into distinct categories based on factors like particle sizes, mineral composition, plasticity, and other engineering characteristics [15–17]. By categorizing soils into classifications like sands, silts, clays, and their combinations, the USCS establishes a uniform terminology for describing soil traits and performance.

Typically, the USCS classification process involves laboratory analyses and field evaluations, where soil samples are scrutinized for parameters like grain size distribution and Atterberg limits, which include plasticity index (PI), plastic limit (PL), and liquid limit (LL), along with other relevant attributes [18–20]. These findings are then compared against predefined classification standards to assign each soil sample to a specific group or category within the classification system. Traditional methods rely heavily on expert knowledge and manual processes, which are time-consuming and prone to human error [21–23]. Machine learning (ML) offers a promising alternative, providing automated and accurate classification based on soil data. Researchers are exploring advanced techniques, including ML algorithms like logistic regression (LR), support vector machine (SVM), random forest (RF), artificial neural networks (ANNs), and decision trees (DTs), to devise effective soil classification methods [24–29]. A concrete example of using ML for soil classification is provided by Aydın et al. [25]. A new classification method for determining different soil classes, based on three ML approaches—support vector classification (SVC), multilayer perceptron (MLP), and RF models—has been proposed by Nguyen et al. [27]. It was indicated by the results that, while all three models perform well, the SVC model is the most accurate in classifying soils. A fuzzy decision tree approach to soil classification is applied by Ribeiro et al. [28]. The application of support vector machines for estimating soil properties and classifying soil types based on known chemical and physical properties in sampled profiles is introduced by Kovačević et al. [29]. The review of the existing literature suggests that ML models hold promise for achieving accurate soil classification tasks.

This study utilizes the RF method, renowned for its high accuracy and robustness, to classify soil types. The parameters for this classification are derived from the USCS, a widely adopted method for categorizing soils based on their properties and behavior. These parameters include grain size distribution, Atterberg limits, coefficient of uniformity (Cu), and coefficient of curvature (Cc), which are defined within specific ranges to create a synthetic database for training the RF model. The importance of input factors in soil classification was evaluated using out-of-bag samples in the RF model. Through rigorous validation techniques, including cross-validation, the performance of the RF model was thoroughly assessed, demonstrating its capability to accurately classify soils. Finally, the RF model developed in this study is utilized to predict the characteristics of 47 soil samples. These predictions are then compared with the actual soil properties to confirm the reliability of the model. Notably, even in scenarios where one input factor is missing, the RF model demonstrates the ability to accurately identify soil properties, a capability that traditional methods lack, thus emphasizing the strengths of the RF model investigated in this study.

## 2. Dataset

The USCS, introduced by Arthur Casagrande, is a widely utilized system for classifying soils based on their properties and behavior, particularly in engineering applications. The primary parameters used in the USCS include grain size distribution, Atterberg limits, soil classification symbols, plasticity chart, Cu, and Cc. These parameters are specified within defined ranges to construct a synthetic database for training the RF model. The parameters compiled in this study are listed in Table 1 below.

**Table 1.** The synthetic database constructed in this study.

	Data Description	Unit	Maximum	Median	Minimum	Standard Deviation	Interval
1	Coefficient of curvature (Cc)	NA	10	4	0	3.42	1
2	Coefficient of uniformity (Cu)	NA	10	5	1	3.18	1
3	Plasticity index (PI)	%	70	35	0	24.15	10
4	Organic soil or inorganic soil	NA	1	0	0	0.50	NA
5	Liquid limit (LL)	%	100	55	0	33.23	10
6	Percentage passing No. 4 sieve	%	100	75	0	27.10	10
7	Percentage passing No. 200 sieve	%	100	35	0	27.94	10
8	Soil classification	NA	25	15	1	8.20	1

Notation: NA is not applicable.

Table 1 presents the descriptions of the datasets used in this study, where Factor 1 to Factor 7 represent the relevant soil property parameters used in USCS evaluation, including Cc, Cu, PI, organic or inorganic soil, LL, percentage passing No. 4 sieve, and percentage passing No. 200 sieve. Factor 8 represents the soil classification. This study establishes a synthetic database to determine the reasonable upper and lower limits for Factor 1 to Factor 7 and then evaluates the soil properties using the USCS. Factor 8 includes 25 types of soil properties. These parameters are defined within specific ranges to build a synthetic database. By establishing the maximum and minimum values of these parameters that are physically meaningful and setting appropriate intervals, as shown in the synthetic database in Table 1, this study has developed a total of 521,316 datasets.

For encoding categorical variables, Factor 4 (organic soil or inorganic soil) differentiates between organic soil and inorganic soil, with 1 indicating organic soil and 0 indicating inorganic soil. Factor 8 relates to soil classification and is encoded from 1 to 25, representing 25 different soil types. The corresponding soil types for each code are summarized in Table 2.

**Table 2.** Definition of soil classification symbol.

No	Soil Type	No	Soil Type	No	Soil Type	No	Soil Type	No	Soil Type
1	CL	6	CH	11	GW-GC	16	GC-GM	21	SP-SM
2	ML	7	MH	12	GP-GM	17	SW	22	SP-SC
3	CL-ML	8	GW	13	GP-GC	18	SP	23	SC
4	OL	9	GP	14	GC	19	SW-SM	24	SM
5	OH	10	GW-GM	15	GM	20	SW-SC	25	SC-SM

Notation: CL is low-plasticity clay, ML is low-plasticity silt, CL-ML is low-plasticity clay and low-plasticity silt; OL is organic silt/clay with low plasticity; OH is organic silt/clay with high plasticity; CH is high-plasticity clay; MH is high-plasticity silt; GW is well graded gravel; GP is poorly graded gravel; GW-GM is well graded gravel and silty gravel; GW-GC is well graded gravel and clayey gravel; GP-GM is poorly graded gravel and silty gravel; GP-GC is poorly graded gravel and clayey gravel; GC is clayey gravel; GM is silty gravel; GC-GM is clayey gravel and silty gravel; SW is well graded sand; SP is poorly graded sand; SW-SM is well graded sand and silty sand; SW-SC is well graded sand and clayey sand; SP-SM is poorly graded sand and silty sand; SP-SC is poorly graded sand and clayey sand; SC is clayey sand; SM is silty sand; SC-SM is clayey sand and silty sand [13,30].

Generally, in soil classification, gravel (G) represents particles larger than 4.75 mm, which are typically coarse and provide good drainage characteristics. Sand (S) represents particles between 0.075 mm and 4.75 mm, offering a range of grain sizes from fine to coarse sand, which influences soil texture and compaction properties. Silt (M) represents particles between 0.002 mm and 0.075 mm, known for their smooth texture and moderate water retention capacity. Clay (C) represents particles smaller than 0.002 mm, characterized by

their fine texture, high plasticity, and significant water retention ability. These classifications help in understanding the soil's mechanical behavior and suitability for various engineering applications.

According to the soil classification symbols, soils are broadly divided into three categories. Coarse-grained soils, which contain more than 50% of particles larger than 0.075 mm, including gravels and sands, known for their good drainage properties and strength. Fine-grained soils, which consist of more than 50% of particles smaller than 0.075 mm, including silts and clays, which are characterized by their plasticity, water retention capacity, and lower permeability. Lastly, highly organic soils are rich in organic matter, often referred to as peat or muck, and are distinguished by their high compressibility and low shear strength. These classifications are essential for understanding the soil's physical and mechanical properties, which influence its behavior and suitability for various engineering applications.

The distribution of the datasets, providing a comprehensive overview of the data, is illustrated in Figure 1. Figure 1 explains that the value ranges for these eight factors encompass the maximum, median, minimum, standard deviation, and interval. This detailed representation allows for a better understanding of the data's spread and variability, highlighting the key statistical measures for each factor. These measures are crucial for analyzing the properties and behavior of soils in accordance with the USCS classification system.

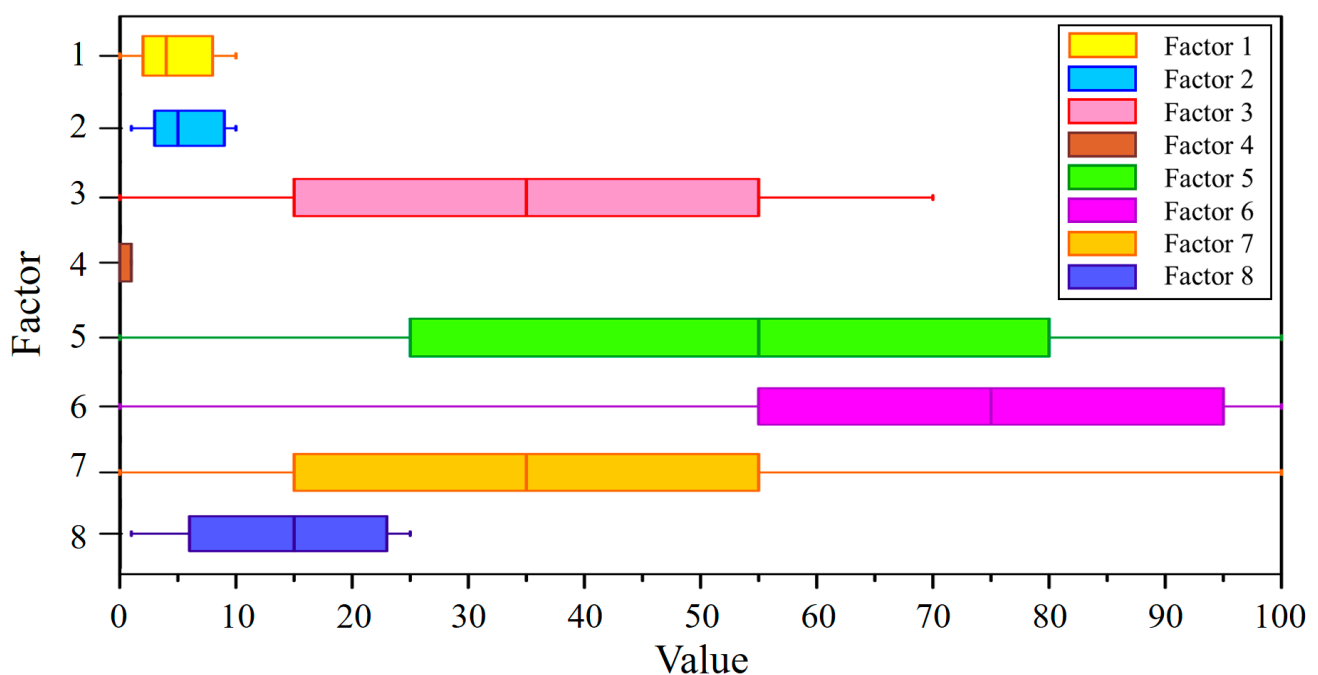


Figure 1. Distribution of datasets.

### 3. Unified Soil Classification System (USCS)

The USCS is commonly employed for the classification of soils according to their physical attributes. This classification method entails several pivotal phases: sample collection and preparation, analysis of grain size distribution, determination of Atterberg limits, classification according to grain size, utilization of the plasticity chart, allocation of group symbols and names, and validation and documentation. The procedure delineated in Figure 2 illustrates the steps involved in the USCS process. Comprehensive explanations of each step are presented in the subsequent sections.

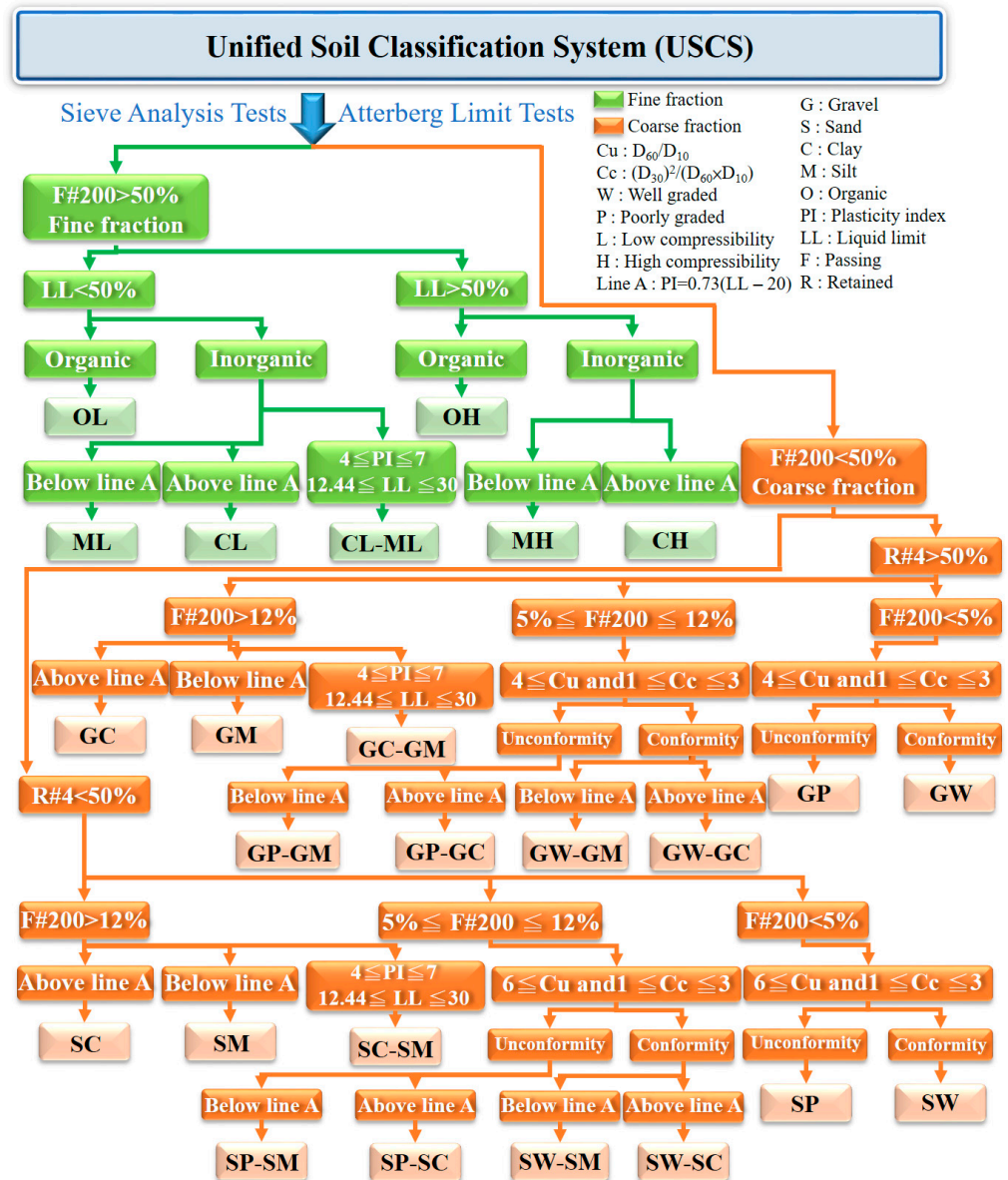


Figure 2. The procedure of USCS.

### 3.1. Sample Collection and Preparation

The initial step is sample collection and preparation. Following the collection of soil samples from the designated site, the subsequent step involves meticulous preparation to ensure accurate analysis. This process commences with the careful drying of the collected samples to eliminate moisture content, thereby preventing any potential alterations in composition during analysis. Additionally, if the collected samples contain larger particles or debris that could impede precise measurements, it is imperative to meticulously sieve them to remove such impediments, ensuring the purity and homogeneity of the samples for further examination.

### 3.2. Grain Size Distribution Analysis

The next phase involves analyzing the distribution of grain sizes, which is crucial for understanding soil composition. This examination initiates with sieve analysis, where the soil is sequentially sieved through a series of screens with decreasing mesh sizes. The quantity of soil retained on each sieve is then meticulously weighed to determine the distribution of grain sizes. Coarse-grained soils, such as sand and gravel, are subjected

to specific sieves like No. 4 (4.75 mm), No. 10 (2.0 mm), No. 40 (0.425 mm), and No. 200 (0.075 mm). Conversely, fine-grained soils like silt and clay undergo hydrometer analysis to accurately assess particle sizes smaller than 0.075 mm. This approach facilitates the precise measurement of the silt and clay fractions' distribution within the soil sample.

In grain size distribution analysis, acquiring parameters like  $C_u$  and  $C_c$  is crucial for comprehending soil properties.  $C_u$  signifies the uniformity of grain sizes within a soil sample and is determined as

$$C_u = \frac{D_{60}}{D_{10}}, \quad (1)$$

where  $D_{60}$  and  $D_{10}$  represent the particle diameters at which 60% and 10% of the sample's mass is finer, respectively. This parameter is crucial as it indicates the soil's gradation: a higher  $C_u$  suggests a well graded soil with a wide range of particle sizes, which generally provides better compaction and stability, while a lower  $C_u$  indicates a poorly graded soil with a narrower range of particle sizes, potentially leading to weaker structural properties.

The  $C_c$  is another important parameter that provides insight into the soil's gradation characteristics. It is calculated as

$$C_c = \frac{(D_{30})^2}{D_{10} \times D_{60}}. \quad (2)$$

In grain size distribution analysis,  $D_{30}$  represents the particle diameter at which 30% of the sample's mass is finer. The  $C_c$  parameter aids in interpreting the shape of the soil's gradation curve, particularly focusing on the curvature around the  $D_{30}$  value. Typically, values between 1 and 3 suggest a well graded soil, while those outside this range indicate a poorly graded soil. This insight is crucial for assessing the soil's appropriateness for engineering and construction endeavors. Well graded soils generally have a superior load-bearing capacity. In terms of drainage, poorly graded soils with uniformly sized particles tend to drain better due to the larger voids between the particles.

### 3.3. Determination of Atterberg Limits

The third step involves the determination of Atterberg limits, which includes measuring the LL, PL, and PI of the soil. These parameters are crucial for assessing the plasticity and behavior of fine-grained soils. The LL represents the moisture content at which the soil transitions from a liquid to a plastic state, while the PL denotes the moisture content at which the soil transitions from a plastic to a semi-solid state. The PI is determined by subtracting the PL from the LL utilizing the following equation:

$$PI = LL - PL. \quad (3)$$

### 3.4. Classification Based on Grain Size

Based on the grain size distribution and Atterberg limits, classify the soil using the USCS chart. Coarse-grained soils (gravel and sand) are classified based on the percentage of fine particles (particles smaller than 0.075 mm) and the grain size distribution. Fine-grained soils (silt and clay) are classified based on their plasticity characteristics and the Atterberg limits. Organic soils are identified if the soil contains a significant amount of organic material, typically indicated by color, odor, and lower specific gravity.

If over 50% of the soil's weight remains on the No. 200 sieve, it falls into the coarse-grained category. Further classification depends on the percentage of sand and gravel present. Soils classified as gravel retain over 50% on the No. 4 sieve. For soils classified as sand, over 50% passes through the No. 4 sieve but is retained by the No. 200 sieve. This classification is based on the coarse fraction. Gravel soils retain more than 50% of their coarse particles on the No. 4 sieve, while sand soils have over 50% of their coarse particles passing through the No. 4 sieve and being retained by the No. 200 sieve. If over 50% of the soil's weight passes through the No. 200 sieve, it is categorized as fine-grained, with sub classification based on the Atterberg limits.



### 3.5. Use of Plasticity Chart

The plasticity chart is a crucial tool in geotechnical engineering, used to classify fine-grained soils based on their LL and PI. This chart aids in distinguishing between silts and clays and assessing their plasticity levels, which can be categorized as low, medium, or high. By graphing the LL and PI on the plasticity chart, engineers can ascertain the soil type and its response under different circumstances. The chart typically has the LL on the *y*-axis and the PI on the *x*-axis, allowing for a clear visualization of the soil's properties.

### 3.6. Assignment of Group Symbols and Names

A two-letter symbol is assigned to the soil based on its classification. For instance, GP signifies poorly graded gravel, SW signifies well graded sand, CL denotes low plasticity clay, and CH indicates high plasticity clay. The USCS offers a systematic and standardized methodology for soil classification, crucial for applications in geotechnical engineering, construction, and other fields requiring an understanding of soil properties and behavior.

For coarse grained soils, the symbols are as follows: GW for well graded gravel, GP for poorly graded gravel, SW for well graded sand, SP for poorly graded sand, GM for silty gravel, GC for clayey gravel, SM for silty sand, and SC for clayey sand. For fine grained soils, the symbols are: ML for low plasticity silt, MH for high plasticity silt, CL for low plasticity clay, CH for high plasticity clay, OL for organic silt/clay with low plasticity, OH for organic silt/clay with high plasticity, and PT for highly organic soils such as Peat.

### 3.7. Reporting

Once the soil group symbol is assigned, the final classification, along with the relevant data, should be documented and reported for engineering and geotechnical applications. The USCS provides a systematic and standardized approach for soil classification, which is essential for geotechnical engineering, construction, and other applications involving soil properties and behavior.

Section 3 thoroughly outlines the USCS procedure, detailing the steps involved in the USCS process. This includes key phases such as sample collection and preparation, grain size distribution analysis, Atterberg limit determination, classification based on grain size, use of the plasticity chart, assignment of group symbols and names, and validation and documentation. This study further uses the random forest method from machine learning to develop an automated soil classification model for the USCS procedure. Detailed explanations of each step are provided in the following section.

## 4. Soil Classification Using Random Forest Method

In this study, a RF model [31] is adopted for the purpose of soil classification. The process of applying the RF methodology to soil classification is depicted in Figure 3. Key to this approach are several critical stages, commencing with the collection and preparation of data. Subsequently, feature selection is carried out to pinpoint the most pertinent attributes for precise classification. After this, the model is built using the chosen features, and hyperparameters are adjusted to enhance its performance. An assessment of the model's effectiveness ensues, followed by thorough testing to ensure its robustness and reliability. In this study, the training data comprise 70% (364,921 data), while the testing data make up 30% (156,395 data). Finally, the results are interpreted and scrutinized to derive meaningful insights. The following sections elaborate on each of these steps in meticulous detail.

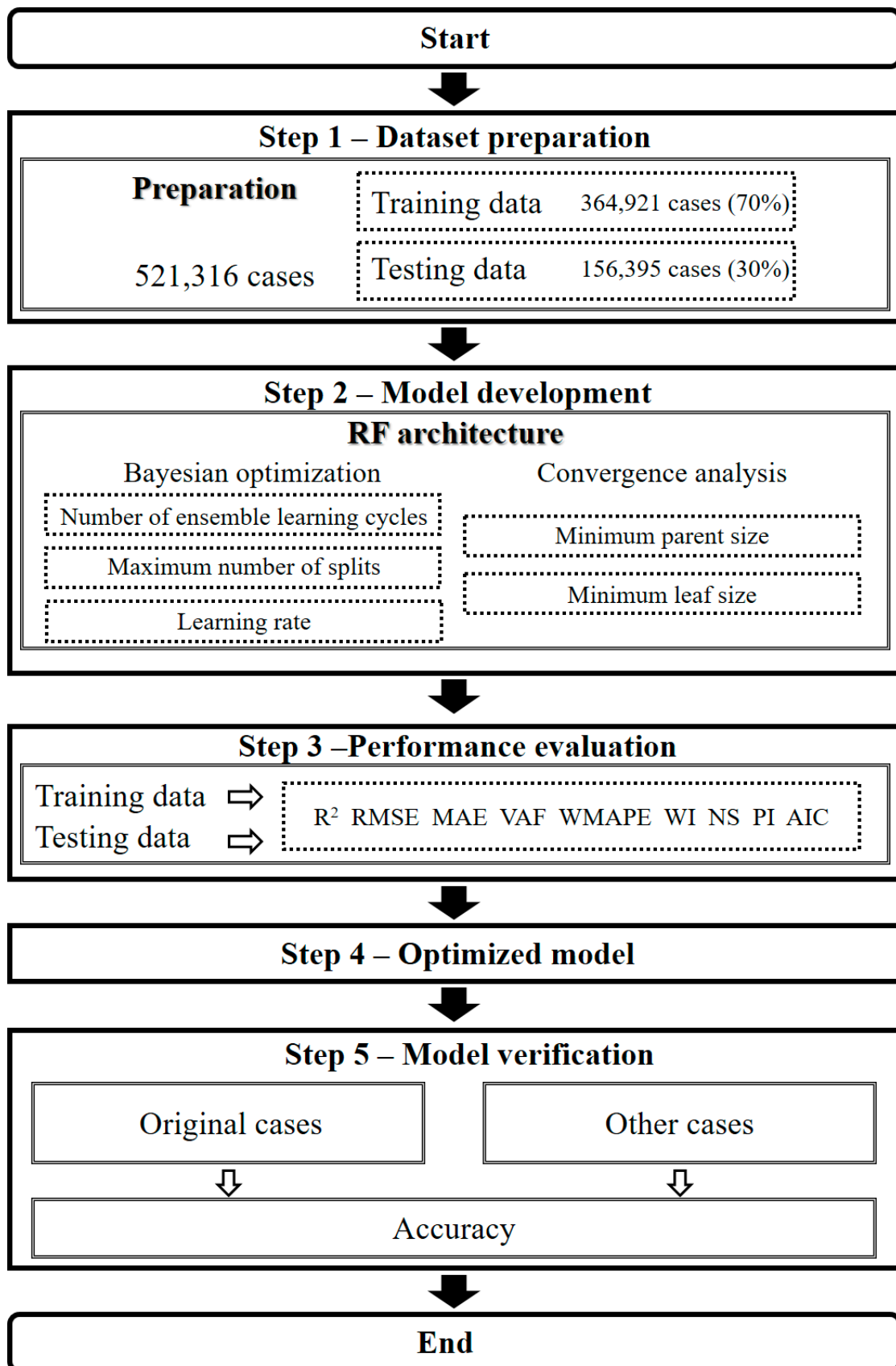


Figure 3. Diagram depicting the RF method for soil classification.



#### 4.1. Data Collection

A comprehensive dataset comprising various features such as Cc, Cu, PI, the classification of soil as organic or inorganic, LL, percentage passing No. 4 sieve, and percentage passing No. 200 sieve, along with their corresponding soil classifications, has been collected for analysis. These parameters are delineated within specified ranges to form a synthetic database for training the RF model. The compiled parameters are detailed in Table 1 provided above. The details of the database established in this study are as follows:

Initially, this study lists the reasonable ranges for seven factors and hypothesizes the possible values of each factor within their respective ranges using different intervals, as listed in Table 1. For example, the Cc has a reasonable range of 0 to 10, and its values are assumed at intervals of 1. Therefore, the values of the Cc in the database are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Other factors are hypothesized in the same way.

Subsequently, different values of each factor are combined, and unreasonable combinations are eliminated. Examples of unreasonable combinations are as follows: (1) The percentage passing the No. 4 sieve must be greater than the percentage passing the No. 200 sieve. Therefore, combinations where the percentage passing the No. 4 sieve is less than the percentage passing the No. 200 sieve are eliminated. (2) The LL of the soil must be greater than the PI. Therefore, combinations where the LL is less than the PI are eliminated. (3) Coarse-grained soils are typically not organic. Therefore, combinations where the soil is both coarse-grained and organic are eliminated.

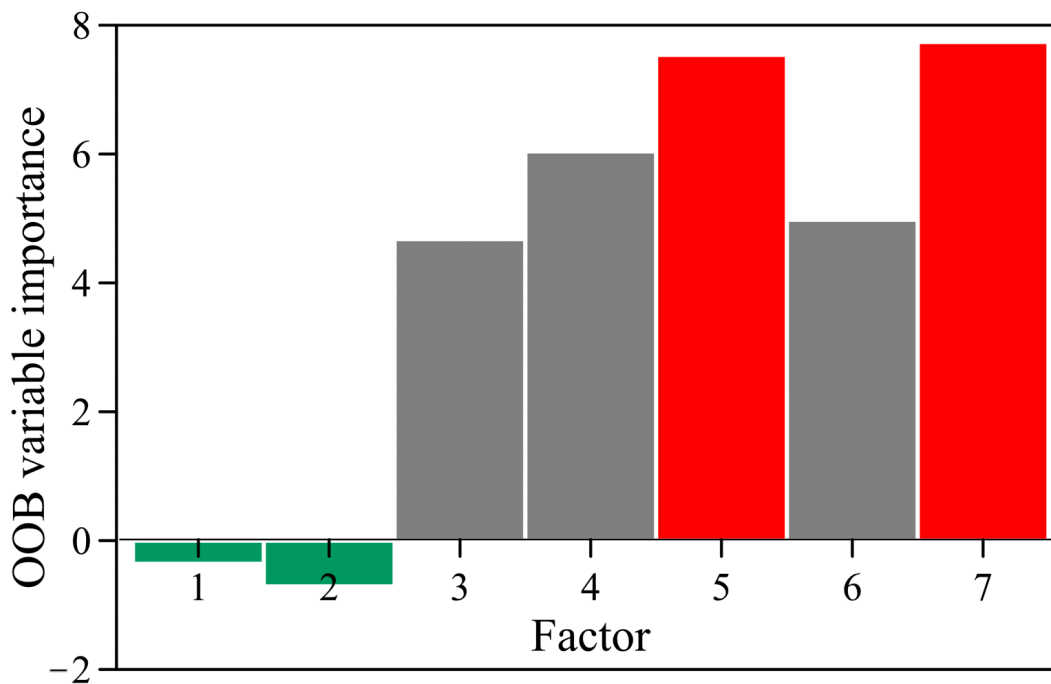
Finally, by integrating the classification criteria for various soils, the results of soil classification are established in Table 1's database. Overall, this study comprises a total of 521,316 datasets. This database is then used for training the RF model.

#### 4.2. Out-of-Bag (OOB) Predictor Importance

In this study, the Out-of-Bag (OOB) feature of the RF model is utilized to assess the importance of seven factors in soil classification. Through the OOB method, we examine the RF model's training process and utilize the OOB samples to estimate the decrease in prediction accuracy when specific feature arrangements are utilized.

The OOB predictor importance serves as a technique for assessing the relevance of features within the RF model. It consists of multiple decision trees, each constructed using distinct training and feature subsets. As these trees are built using different random samples and features, certain data points may never be utilized during training, known as OOB data. The OOB predictor importance evaluates each feature's contribution to the model based on its performance in the OOB testing. This importance is determined by the frequency of feature usage across all trees and the average reduction in testing error observed when the feature is employed in each tree. The OOB predictor importance offers an intuitive means to gauge features' impact on the RF model's predictive performance and aids in selecting the most crucial features for modeling purposes. By integrating the importance of the OOB predictors, influential predictors are identified within the RF model utilized.

The analysis findings regarding variable importance using the OOB predictor, as shown in Figure 4, reveal intriguing insights. Factors 1 (Cc) and 2 (Cu) exhibit negative importance values, indicating their diminished relevance in the classification process. The OOB analysis results of this study align well with the facts. Specifically, the coefficients Cc and Cu are not employed for classifying all soil types. These coefficients are primarily used to describe the grain size distribution of coarse-grained soils. They are not applicable to fine-grained soils (such as silt and clay), which lack a significant range of particle sizes, and are also not used for coarse soils with fines greater than 12% or for organic soils. Conversely, Factors 5 (LL) and 7 (percentage passing No. 200 sieve) demonstrate the highest importance values, highlighting their pivotal roles in soil classification. This observation implies that LL and the percentage passing No. 200 sieve are not only critical but also integral factors that significantly influence the outcome of the classification process, emphasizing the need for careful consideration in soil analysis and interpretation.



**Figure 4.** Variable importance plot using the OOB predictor (green indicates that the OOB variable importance is negative, gray represents an OOB variable importance between 0 and 6, and red signifies an OOB variable importance greater than 6).

#### 4.3. Model Construction

The process of initializing the RF model begins with a predetermined number of decision trees. Subsequently, it undergoes training on the designated dataset, where each tree's growth occurs through the utilization of a bootstrap sample of the data. At each node, the model selects the optimal split based on a subset of features. Illustrated in Figure 5, the RF's architecture involves the assembly of multiple decision trees, with each tree constructed independently utilizing a random subset of the training data and input features. For this study, the training and testing datasets constitute 70% and 30% of the total data, respectively. Throughout the training phase, each tree expands either until it reaches its maximum depth or meets a specified stopping criterion, such as the minimum number of samples required for node splitting or a maximum depth threshold.

Once all the decision trees are built, predictions are produced by consolidating the outputs from each individual tree. In regression tasks, the final prediction is typically the average of all tree predictions, while in classification tasks, it is usually determined by a majority vote among the trees. The inclusion of randomness in the construction of each tree aids in diminishing correlations among the trees and mitigating the risk of overfitting. Assuming there exists a database  $D$ , it can be represented as follows:

$$(x_i, y_i) \text{ for } i = 1, 2, \dots, N \ \& \ x_{i1}, x_{i2}, \dots, x_{ip}, \quad (4)$$

In this equation,  $x$ ,  $y$ ,  $N$ , and  $p$  are the input, output, data number, and number of factors. If  $D$  is divided into  $M$  regions and  $D_1, D_2, \dots, D_M$  is obtained, and a constant  $c_m$  is used to represent the simulated output  $f(x)$  of each region, the following equation can be obtained:

$$f(x) = \sum_{m=1}^M c_m I(x \in D_m), \quad (5)$$

where  $I$  is an indicator function. By incorporating the least squares sum as a criterion, the optimal constant,  $\hat{c}_m$ , can be obtained as the average of the output values,  $D_m$ , within the region:

$$\hat{c}_m = \text{average} (y_i | x_i \in D_m). \tag{6}$$

Assuming the presence of a categorical variable  $j$  and a designated split point  $s$ , the database is partitioned into two distinct subsets, as indicated by the following equation:

$$D_1(j, s) = \{x | x_j \leq s\} \text{ and } D_2(j, s) = \{x | x_j > s\}. \tag{7}$$

As per the preceding equation, the quest for the suitable categorical variable  $j$  and split point  $s$  results in the following equation:

$$\min_{j,s} \left[ \min_{c_1} \sum_{x_i \in D_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in D_2(j,s)} (y_i - c_2)^2 \right]. \tag{8}$$

Referring to the equation above, the internal minimization for any combination of  $j$  and  $s$  can be deduced from the subsequent expression:

$$\hat{c}_1 = \text{average} (y_i | x_i \in D_1(j, s)) \text{ \& } \hat{c}_2 = \text{average} (y_i | x_i \in D_2(j, s)). \tag{9}$$

Utilizing the aforementioned equations, the optimal pair  $(j, s)$  can be determined, facilitating the partitioning of the data into two regions. Iterating through the described computations enables the data to be sequentially split into all resulting regions. If a decision tree  $T$  partitions the data into  $D_m$  regions via  $m$  nodes, where  $N_m$  represents the total number of regions,  $\hat{c}_m$  can be articulated as follows:

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in D_m} y_i. \tag{10}$$

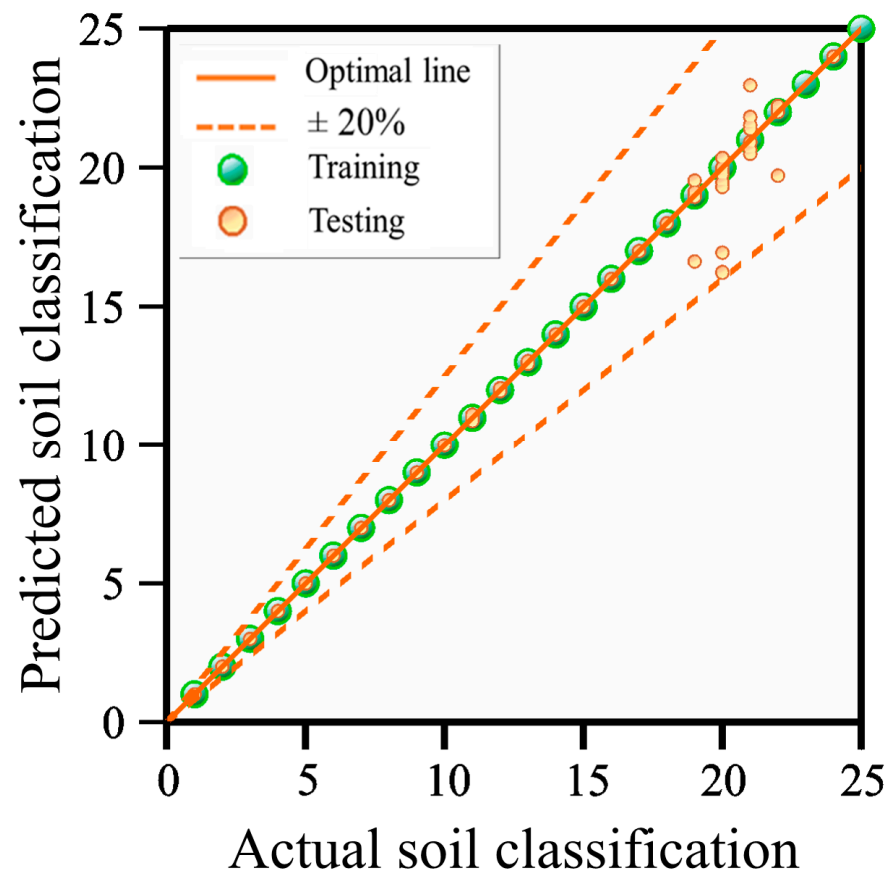
Bagging, or bootstrap aggregation, is a technique employed to acquire an aggregated predictor by creating numerous predictor variations and amalgamating them. When this aggregated predictor is employed for numerical prediction, it calculates the average of the results from each variation and may also conduct a majority vote on prediction outcomes. Different predictor variations are obtained by sampling from the dataset, with each sampling akin to modeling a novel dataset.

Assuming a database  $D$  as described earlier, it is divided into smaller datasets,  $\tilde{D}_b$ , and  $b = 1, 2, \dots, B$  to obtain  $\tilde{D}_1, \tilde{D}_2, \dots, \tilde{D}_B$ . The sampling process involves a fixed number of samples each time, and the sampled data are replaced back into the original dataset before the next sampling. After calculating each small dataset,  $\tilde{D}_b$ , using the base algorithm, their results,  $\tilde{f}(x)$ , are collected, and the final training result is obtained by averaging all results,  $\tilde{f}_{bag}(x)$ , expressed as follows:

$$\tilde{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \tilde{f}_b(x), \tag{11}$$

where  $\tilde{f}_b(x)$  is the output results obtained for each small dataset by the base algorithm.

In an RF model, each decision tree necessitates the configuration of specific hyperparameters, including the learning rate, number of iterations, and number of features employed for node partitioning, among others. These hyperparameters' selection significantly influences the model's performance. In this study, the proposed model is trained on bootstrap samples from the dataset, and predictions are aggregated to produce the final output. While the RF method naturally utilizes bagging, assessing the performance of the ensemble itself offers valuable insights into model stability and accuracy.



**Figure 5.** Performance of proposed RF model.

#### 4.4. Hyperparameter Tuning

Critical hyperparameters like the number of trees, maximum tree depth, and the number of features assessed for splitting at each node are carefully determined. This thorough analysis involves evaluating the convergence of the model's performance across a range of hyperparameters to enhance its robustness and efficiency. By examining the model's performance under different hyperparameter configurations, we can identify the most optimal settings to attain superior results.

#### 4.5. Model Performance Evaluation

The RF model undergoes evaluation using a comprehensive range of performance indices to ensure a thorough assessment of its accuracy and effectiveness. These performance indices include the variance accounted for (VAF), coefficient of determination ( $R^2$ ), prediction interval (PI), Nash–Sutcliffe efficiency (NS), weighted index (WI), Akaike information criterion (AIC), root mean square error (RMSE), weighted mean absolute percentage error (WMAPE), and mean absolute error (MAE) [32]. Each of these indices plays a crucial role in evaluating different aspects of the model's performance, from its ability to explain variance to its predictive precision and overall fit to the data. By collectively calculating scores across these performance indices, the evaluation process provides a robust validation of the RF model's accuracy and reliability in soil classification tasks.

#### 4.6. Model Validation

Cross-validation ensures the model's performance across various data subsets, validating its ability to generalize to unseen data. This method systematically divides the dataset into multiple subsets, training the model on some subsets while testing it on others. By repeating this process iteratively, cross-validation offers a comprehensive evaluation of the RF method's predictive capabilities. The reliability of the predictions is evaluated

using various performance indices. These indices provide a detailed understanding of the model's strengths and weaknesses, ensuring robust and accurate soil classification. By analyzing these metrics, the study aims to fine-tune the model for optimal performance, thereby enhancing the reliability of soil classification predictions.

## 5. Results and Discussion

In this section, we outline the verification and application scenarios addressed in our study. During the verification phase, we initially develop and validate models using the RF algorithm, incorporating the relevant factors as inputs. However, when real-life constraints such as time or budget limitations make it difficult to obtain comprehensive geological data, only partial input factors are considered. In these instances, soil classification evaluations are performed using the RF model developed in this study.

### 5.1. Performance of the Proposed RF

In the validation cases, this study uses seven factors: Cc, Cu, PI, organic or inorganic soil classification, LL, percentage passing No. 4 sieve, and percentage passing No. 200 sieve. These factors are selected based on their reasonable ranges and their relevance to accurately identifying soil properties. The training data constitute 70% (364,921 data), while the testing data account for 30% (156,395 data).

Subsequently, this study employs the RF model for soil classification. The parameters used in this RF model include the following: the number of variables to sample is set to all, surrogate splits are enabled, and pruning is disabled. The splitting criterion used is mean squared error (MSE), and the type of model is regression. The quadratic error tolerance is set to  $10^{-6}$ , with a minimum parent size of 10 and a minimum leaf size of 1. The learning rate, maximum number of splits, and number of ensemble learning cycles are all determined using Bayesian optimization [33–35]. These parameters collectively define the configuration of the RF model used in this study.

Figure 5 provides a visual representation of the correlation coefficient computed specifically for the training dataset. Upon examination, it becomes evident that both the training and testing datasets produce remarkably accurate results, exhibiting a notably high correlation coefficient of 0.99. This implies a robust alignment between the predicted values and the actual observations. As the data points closely conform to the optimal line, it signifies enhanced precision in classifying the soil properties. Furthermore, the validation cases corroborate these findings, revealing a compelling relationship between the soil property predictions derived from the seven factors outlined in this study and the real soil properties. This strong correlation underscores the efficacy and reliability of the analytical framework employed for soil characterization.

Due to the inherent stochastic elements in the ML models and the influence of training data, the model's performance can exhibit variability across different training sets, resulting in slight disparities in results with each computation. To address this, in the present study, the ML model underwent 50 iterations, and the average values of relevant performance indices were computed to provide a more robust assessment. The analysis findings, outlined in Table 3, illustrate that the VAF,  $R^2$ , PI, NS, WI, AIC, RMSE, MAE, and WMAPE are 100, 1, 2, 1, 1,  $-2.23 \times 10^7$ ,  $7.96 \times 10^{-4}$ ,  $6.48 \times 10^{-6}$ , and  $1.16 \times 10^{-5}$ , respectively. These analysis outcomes underscore the robustness of the proposed RF model in effectively discerning the characteristics of soil, providing valuable insights into soil classification and prediction.

**Table 3.** The performance of the proposed RF across 50 runs.

Performance Indices	Ideal Value	Training	Testing
VAF	100	100	100
$R^2$	1	1	1
PI	2	2	2

Table 3. Cont.

Performance Indices	Ideal Value	Training	Testing
NS	1	1	1
WI	1	1	1
AIC	NA	$-1.18 \times 10^8$	$-2.24 \times 10^7$
RMSE	0	$9.85 \times 10^{-8}$	$7.96 \times 10^{-4}$
MAE	0	$4.02 \times 10^{-9}$	$6.48 \times 10^{-6}$
WMAPE	0	$7.19 \times 10^{-9}$	$1.16 \times 10^{-5}$

### 5.2. Validation

Conventional soil classification methods rely on a thorough examination of various soil parameters, including coefficients of curvature and uniformity, PI, organic or inorganic soil classification, LL, percentage passing No. 4 sieve, and percentage passing No. 200 sieve, totaling seven factors. However, practical limitations such as time constraints, budgetary restrictions, or data availability may impede the acquisition of all seven parameters. Therefore, this study investigates whether the RF model can effectively evaluate soil properties even when some input soil parameters are unavailable. Its aim is to assess the model's ability to compute performance indices and accurately identify soil properties in such scenarios.

This study encompasses eight cases, where Case 1 corresponds to scenarios with no missing factors, indicating the utilization of the complete set of seven factors as input variables. Cases 2 to 8, conversely, involve the absence of one factor each. A summary detailing the missing factors for each case is presented in Table 4. In this study, nine performance indices, VAF,  $R^2$ , PI, NS, WI, AIC, RMSE, MAE, and WMAPE values, were employed to calculate the scores, with the highest accuracy cases receiving the highest scores and vice versa. This study initially calculated nine performance indices for each of the eight cases. The performance index values for the eight cases are summarized in Table 4.

Table 4. Performance index values across eight cases.

Case	Missing Data	$R^2$	RMSE	VAF	PI	MAE	WI	WMAPE	NS	AIC
1	NA	1	$7.96 \times 10^{-4}$	100	2.00	$6.48 \times 10^{-6}$	1	$1.16 \times 10^{-5}$	1	$-2.24 \times 10^7$
2	Factor 1	1	$6.36 \times 10^{-3}$	99.97	1.99	$1.46 \times 10^{-3}$	1	$2.63 \times 10^{-3}$	1	$-1.58 \times 10^7$
3	Factor 2	1	$4.41 \times 10^{-3}$	99.98	2.00	$7.03 \times 10^{-4}$	1	$1.27 \times 10^{-3}$	1	$-1.70 \times 10^7$
4	Factor 3	0.99	$3.14 \times 10^{-2}$	99.15	1.95	$2.55 \times 10^{-2}$	1	$4.58 \times 10^{-2}$	0.99	$-1.08 \times 10^7$
5	Factor 4	0.99	$2.97 \times 10^{-2}$	99.24	1.96	$2.10 \times 10^{-2}$	1	$3.79 \times 10^{-2}$	0.99	$-1.10 \times 10^7$
6	Factor 5	0.98	$4.67 \times 10^{-2}$	98.1	1.92	$2.62 \times 10^{-2}$	1	$4.73 \times 10^{-2}$	0.98	$-9.58 \times 10^6$
7	Factor 6	0.82	$1.45 \times 10^{-1}$	77.79	1.45	$1.21 \times 10^{-1}$	0.95	$2.18 \times 10^{-1}$	0.78	$-6.03 \times 10^6$
8	Factor 7	0.09	$3.26 \times 10^{-1}$	0.69	0.77	$2.83 \times 10^{-1}$	0.4	$5.11 \times 10^{-1}$	0.54	$-3.50 \times 10^6$

Notation: Factor 1 is Cc, Factor 2 is Cu, Factor 3 is PI, Factor 4 is organic soil or inorganic soil, Factor 5 is LL, Factor 6 is percentage passing No. 4 sieve, and Factor 7 is percentage passing No. 200 sieve.

This study further calculates scores for these eight cases based on nine performance indices, ranks them, and assigns scores to each case. Table 5 lists the results of scores across eight cases. The calculation of scores is explained as follows: scores ranged from eight points for the top-performing case (such as Case 1, achieving the highest  $R^2$  score) to eight points for the least-performing case (Case 8). The remaining cases were scored in descending order according to their respective error metric results. Each case was evaluated



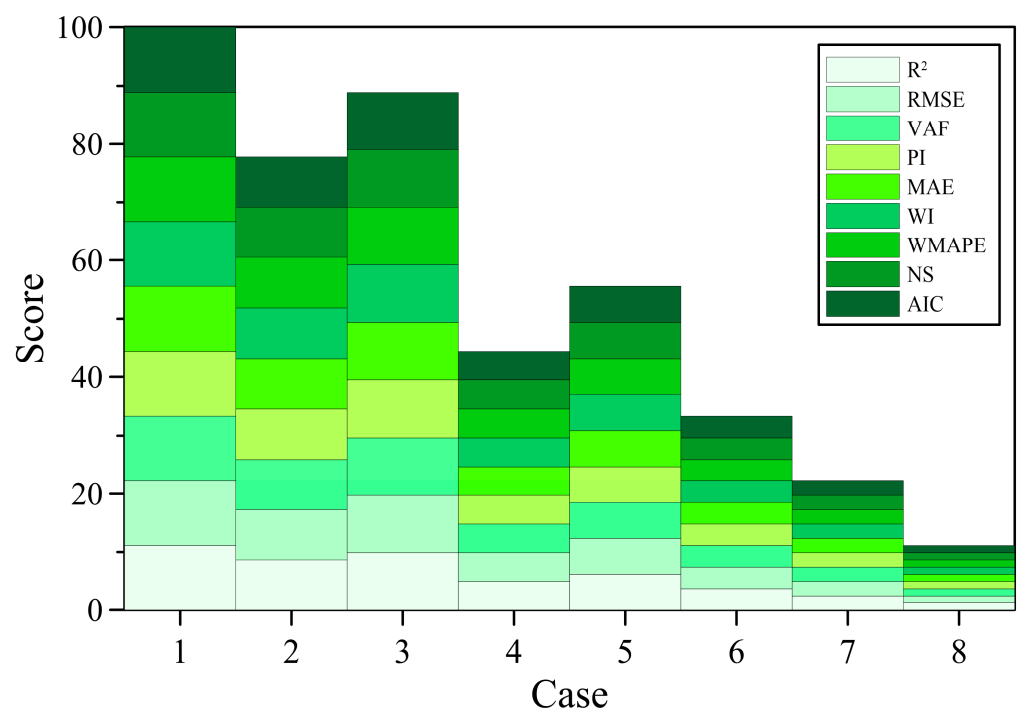
across all nine performance indices, and these scores were summed to derive a total sum. In Case 1, the maximum total sum (sum = 72) was obtained.

**Table 5.** Results of scores across eight cases.

Case	R <sup>2</sup>	RMSE	VAF	PI	MAE	WI	WMAPE	NS	AIC	Sum	Score
1	8	8	8	8	8	8	8	8	8	72	100
2	6	6	6	6	6	6	6	6	6	54	77.78
3	7	7	7	7	7	7	7	7	7	63	88.89
4	4	4	4	4	4	4	4	4	4	36	44.44
5	5	5	5	5	5	5	5	5	5	45	55.56
6	3	3	3	3	3	3	3	3	3	27	33.33
7	2	2	2	2	2	2	2	2	2	18	22.22
8	1	1	1	1	1	1	1	1	1	9	11.11

Subsequently, this sum was converted into a score using the following formula: the score is equal to the sum divided by the maximum total sum across eight cases (maximum total sum = 72) and then multiplied by 100. This transformation aids in comparing and presenting scores relative to different standards, making comparisons more intuitive. The scores for the nine performance indices across the eight cases are summarized.

Finally, the scores for the nine performance indices across the eight cases are obtained, as listed in Table 5. An accuracy assessment is conducted for each metric, with the case achieving the highest precision for each metric receiving the top score. The scores are thus aggregated, with the maximum attainable score being 100. The score outcomes for the eight cases investigated are depicted in Figure 6. The analysis highlights that Case 1 emerged with the highest score, showcasing its superior performance. Similarly, Cases 2 and 3 also achieved commendable scores, surpassing the 70 score. Conversely, the remaining cases fell short, scoring below 60.



**Figure 6.** Results of nine performance indices for eight cases.

Typically, traditional USCS analysis relies on utilizing seven factors to accurately classify soil. However, obtaining comprehensive parameters directly on-site may not always be feasible in real-world situations. Therefore, the model formulated in this study has demonstrated its effectiveness in classifying soil accurately, even when confronted with incomplete input factors. This capability to achieve precise soil classification with limited input factors distinguishes the developed model from conventional methods, highlighting its strength and adaptability in navigating the practical constraints encountered in soil classification tasks.

### 5.3. Prediction

Based on the previous validation section, Case 1, Case 2, and Case 3 achieved commendable scores, each exceeding 70, indicating relatively reliable models. The remaining cases scored below 60, making those models relatively unreliable. Therefore, this study further evaluates the models using only these three cases, with 47 different soil samples as input factors. Table 6 lists the datasets of 47 soil samples [36,37]. As listed in Table 6, an extensive dataset comprising 47 different soil samples was consolidated, and each type was assigned a unique identifier ranging from 1 to 47. The data sources for samples 1 through 15 are derived from Das et al. [36], while samples 16 through 41 are sourced from Das and Sobhan [37]. Samples 42 through 47 are based on the soil mechanic experimental data obtained in this study.

**Table 6.** Datasets of 47 soil samples [36,37].

Soil Sample	Coefficient of Curvature (Cc)	Coefficient of Uniformity (Cu)	Plasticity Index (PI)	Organic Soil or Inorganic Soil	Liquid Limit (LL)	Percentage Passing No. 4 Sieve	Percentage Passing No. 200 Sieve	Soil Classification
1	0	1	10	0	30	100	58	CL
2	0	1	12	0	33	70	30	SC
3	0	1	21	0	33	70	30	SC
4	0	1	22	0	41	48	20	GC
5	0	1	28	0	52	95	70	CH
6	0	1	19	0	30	100	82	CL
7	0	1	21	0	35	100	74	CL
8	0	1	18	0	38	87	26	SC
9	0	1	38	0	69	88	78	CH
10	0	1	26	0	54	99	57	CH
11	4.8	2.9	16	0	32	71	11	SP-SC
12	7.2	2.2	0	0	0	100	2	SP
13	0	1	21	0	44	89	65	CL
14	3.9	2.1	31	0	39	90	8	SP-SC
15	0	1	4	0	23	100	13	SC-SM
16	0	1	25	0	63	100	77	MH
17	1.59	3.44	0	0	0	94	3	SP
18	0	1	22	0	37	100	65	CL
19	0	1	21	0	40	100	63	CL
20	0	1	4	0	23	100	13	SC-SM
21	1.59	3.44	0	0	0	94	3	SP
22	0	1	25	0	63	100	77	MH
23	0	1	28	0	55	100	86	CH
24	0	1	22	0	36	100	45	SC
25	0	1	8	0	30	92	48	SC

Table 6. Cont.

Soil Sample	Coefficient of Curvature (Cc)	Coefficient of Uniformity (Cu)	Plasticity Index (PI)	Organic Soil or Inorganic Soil	Liquid Limit (LL)	Percentage Passing No. 4 Sieve	Percentage Passing No. 200 Sieve	Soil Classification
26	0	1	4	0	26	60	40	SC-SM
27	0	1	32	0	60	99	76	CH
28	0	1	12	0	41	90	60	ML
29	0	1	2	0	24	80	35	SM
30	0	1	21	0	33	70	30	SC
31	0	1	22	0	41	48	20	GC
32	0	1	28	0	52	95	70	CH
33	0	1	19	0	30	100	82	CL
34	0	1	21	0	35	100	74	CL
35	0	1	18	0	38	87	26	SC
36	0	1	38	0	69	88	78	CH
37	0	1	26	0	54	99	57	CH
38	4.8	2.9	16	0	32	71	11	SP-SC
39	7.2	2.2	0	0	0	100	2	SP
40	0	1	21	0	44	89	65	CL
41	3.9	2.1	31	0	39	90	8	SP-SC
42	0.73	10	3.49	0	23.93	84.7	10.96	SP-SC
43	0.28	10	3.17	0	22.38	70	10	SP-SC
44	0.65	9.81	3.16	0	25.81	71.32	9.18	SP-SM
45	0.22	10	1.33	0	24.76	70.37	9.23	SP-SM
46	0.59	10	1.33	0	24.76	81	13	SM
47	0.52	10	2.496	0	24.328	69.63	6.99	SP-SM

The proposed RF model was employed for analysis and soil classification, with the predicted soil properties compared against actual properties. The RF model's accuracy results for the 47 soil samples across these cases are illustrated in Table 7. The obtained results show that 46 of 47 samples are correctly predicted when none of the factors are missing (Case 1). When Factor 1 (Cc) is missing, the predicted results show that 44 out of 47 samples are correctly predicted. When Factor 2 (Cu) is absent, the predicted results indicate that 45 out of 47 samples are correctly classified. These predicted results support the OOB importance analysis using the proposed RF method. Factor 1 (Cc) and Factor 2 (Cu) exhibited lower importance, thus their absence did not significantly affect the model performance.

Table 7. Predicted results for 47 soil samples.

Case	Data	Predicted Results
1	All data included	46 of 47 samples are correctly predicted
2	Cc is neglected	44 of 47 samples are correctly predicted
3	Cu is neglected	45 of 47 samples are correctly predicted

## 6. Conclusions

This study uses an ML model, particularly RF, to classify soil. Through this innovative approach, the developed RF model effectively classifies soil. The main findings of this study can be summarized as follows:

- (1) This study first employs the OOB predictor within the RF algorithm to evaluate variable importance. The analysis reveals that Factor 5 (LL) and Factor 7 (percentage passing No. 200 sieve) possess the highest importance scores. Since the key components of the USCS are grain size distribution and Atterberg limits, with Factor 5 (LL) and Factor 7 (percentage passing No. 200 sieve) being crucial at the first and second levels of classification, the results imply that they significantly influence the classification outcome.
- (2) Through cross-validation, the performance of the RF model is rigorously evaluated, demonstrating its capability for accurate soil classification. The indices illustrate that the VAF,  $R^2$ , PI, NS, WI, AIC, RMSE, MAE, and WMAPE values are 100, 1, 2, 1, 1,  $-2.23 \times 10^7$ ,  $7.96 \times 10^{-4}$ ,  $6.48 \times 10^{-6}$ , and  $1.16 \times 10^{-5}$ , respectively. These results highlight the robustness and reliability of the proposed RF model in effectively discerning soil characteristics.
- (3) Furthermore, this study examines the impact of missing input factors on the RF model's ability to classify soil characteristics. The analysis reveals that this trend aligns with the findings from the OOB importance analysis. The results demonstrate that the USCS is already optimized regarding the laboratory work. Since Factor 1 (Cc) and Factor 2 (Cu) are calculated from the grain size distribution, the proposed RF model still achieved accurate soil classification despite the omission of these factors.

**Author Contributions:** Methodology, investigation and writing, C.-Y.L.; conceptualization and supervision, C.-Y.K.; validation and visualization, T.-Y.W.; data curation, Y.-C.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Hartemink, A.E. The use of soil classification in journal papers between 1975 and 2014. *Geoderma Reg.* **2015**, *5*, 127–139. [[CrossRef](#)]
2. Bhattacharya, B.; Solomatine, D. Machine learning in soil classification. *Neural Netw.* **2006**, *19*, 186–195. [[CrossRef](#)] [[PubMed](#)]
3. da Silva, A.F.; Pereira, M.J.; Carneiro, J.D.; Zimback, C.R.L.; Landim, P.M.B.; Soares, A. A new approach to soil classification mapping based on the spatial distribution of soil properties. *Geoderma* **2014**, *219–220*, 106–116. [[CrossRef](#)]
4. Elbasi, E.; Zaki, C.; Topcu, A.E.; Abdelbaki, W.; Zreikat, A.I.; Cina, E.; Shdefat, A.; Saker, L. Crop prediction model using machine learning algorithms. *Appl. Sci.* **2023**, *13*, 9288. [[CrossRef](#)]
5. Obasi, S.N.N.; Pemberton, J.; Awe, O.O. A comparative study of soil classification machine learning models for construction management. *Int. J. Constr. Manag.* **2024**, 1–10. [[CrossRef](#)]
6. Xiang, M.; Li, Y.; Yang, J.; Li, Y.; Li, F.; Hu, B.; Cao, Y. Assessment of heavy metal pollution in soil and classification of pollution risk management and control zones in the industrial developed city. *Environ. Manag.* **2020**, *66*, 1105–1119. [[CrossRef](#)] [[PubMed](#)]
7. Yusnita, D.; Ahmad, A.; Solle, M.S. Soil classification for sustainable agriculture. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2020; Volume 486, p. 012045.
8. Avery, B.W. Soil classification in the Soil Survey of England and Wales. *Eur. J. Soil Sci.* **1973**, *24*, 324–338. [[CrossRef](#)]
9. Canadian Agricultural Services Coordinating Committee; Soil Classification Working Group; National Research Council Canada, Canada; Agriculture, & Agri-Food Canada, Research Branch. *The Canadian System of Soil Classification (No. 1646)*; NRC Research Press: Ottawa, ON, Canada, 1998.
10. Abdelfattah, M.A.; Shahid, S.A. A comparative characterization and classification of soils in Abu Dhabi coastal area in relation to arid and semi-arid conditions using USDA and FAO soil classification systems. *Arid. Land Res. Manag.* **2007**, *21*, 245–271. [[CrossRef](#)]
11. Hempel, J.; Micheli, E.; Owens, P.; McBratney, A. Universal soil classification system report from the International Union of Soil Sciences Working Group. *Soil Horizons* **2013**, *54*, 1–6. [[CrossRef](#)]
12. Michéli, E.; Láng, V.; Owens, P.R.; McBratney, A.; Hempel, J. Testing the pedometric evaluation of taxonomic units on soil taxonomy—A step in advancing towards a universal soil classification system. *Geoderma* **2016**, *264*, 340–349. [[CrossRef](#)]

13. Alade, S.M. Correlation of unified and AASHTO soil classification systems for soils classification. *J. Earth Sci. Geotech. Eng.* **2018**, *8*, 39–50.
14. Soltani, A.; O’Kelly, B.C. Reappraisal of the ASTM/AASHTO standard rolling device method for plastic limit de-termination of fine-grained soils. *Geosciences* **2021**, *11*, 247. [[CrossRef](#)]
15. Park, J.; Santamarina, J.C. Revised soil classification system for coarse-fine mixtures. *J. Geotech. Geoenviron. Eng.* **2017**, *143*, 04017039. [[CrossRef](#)]
16. Gadouri, H.; Harichane, K.; Ghrici, M. Assessment of sulphates effect on the classification of soil–lime–natural pozzolana mixtures based on the Unified Soil Classification System (USCS). *Int. J. Geotech. Eng.* **2018**, *12*, 293–301. [[CrossRef](#)]
17. Cook, R.; Fox, T.R.; Allen, H.L.; Cohrs, C.W.; Ribas-Costa, V.; Trlica, A.; Ricker, M.; Carter, D.R.; Rubilar, R.; Campoe, O.; et al. Forest soil classification for intensive pine plantation management: “Site Productivity Optimization for Trees” system. *For. Ecol. Manag.* **2024**, *556*, 121732. [[CrossRef](#)]
18. Warren, S.N.; Kallu, R.R.; Barnard, C.K. Correlation of the rock mass rating (RMR) system with the unified soil classification system (USCS): Introduction of the weak rock mass rating system (W-RMR). *Rock Mech. Rock Eng.* **2016**, *49*, 4507–4518. [[CrossRef](#)]
19. El Majid, A.; Cherradi, C.; Khadija, B.A.B.A.; Razzouk, Y. Laboratory investigations on the behavior of CBR in two expanding soils reinforced with plant fibers of varying lengths and content. *Mater. Today Proc.* **2023**; *in press*. [[CrossRef](#)]
20. Eslami, A.; Golafzani, S.H.; Naghibi, M.H. Developed triangular charts; deltaic CPTu-based soil behavior classification using AUT: CPTu-Geo-Marine Database. *Probab. Eng. Mech.* **2023**, *71*, 103380. [[CrossRef](#)]
21. Padarian, J.; Minasny, B.; McBratney, A. Using deep learning to predict soil properties from regional spectral data. *Geoderma Reg.* **2019**, *16*, e00198. [[CrossRef](#)]
22. Chala, A.T.; Ray, R. Assessing the performance of machine learning algorithms for soil classification using cone penetration test data. *Appl. Sci.* **2023**, *13*, 5758. [[CrossRef](#)]
23. Abraham, S.; Huynh, C.; Vu, H. Classification of Soils into Hydrologic Groups Using Machine Learning. *Data* **2020**, *5*, 2. [[CrossRef](#)]
24. Gambill, D.R.; Wall, W.A.; Fulton, A.J.; Howard, H.R. Predicting USCS soil classification from soil property variables using Random Forest. *J. Terramechanics* **2016**, *65*, 85–92. [[CrossRef](#)]
25. Aydın, Y.; Işıkdag, Ü.; Bekdaş, G.; Nigdeli, S.M.; Geem, Z.W. Use of machine learning techniques in soil classification. *Sustainability* **2023**, *15*, 2374. [[CrossRef](#)]
26. Pham, B.T.; Nguyen, M.D.; Nguyen-Thoi, T.; Ho, L.S.; Koopialipoor, M.; Quoc, N.K.; Armaghani, D.J.; Van Le, H. A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. *Transp. Geotech.* **2021**, *27*, 100508. [[CrossRef](#)]
27. Nguyen, M.D.; Costache, R.; Sy, A.H.; Ahmadzadeh, H.; Van Le, H.; Prakash, I.; Pham, B.T. Novel approach for soil classification using machine learning methods. *Bull. Eng. Geol. Environ.* **2022**, *81*, 468. [[CrossRef](#)]
28. Ribeiro, M.V.; Cunha LM, S.; Camargo, H.A.; Rodrigues, L.H.A. Applying a fuzzy decision tree approach to soil classification. In Proceedings of the Information Processing and Management of Uncertainty in Knowledge-Based Systems: 15th International Conference, IPMU 2014, Montpellier, France, 15–19 July 2014; Proceedings, Part I 15. Springer: Cham, Switzerland, 2014; pp. 87–96.
29. Kovačević, M.; Bajat, B.; Gajić, B. Soil type classification and estimation of soil properties using support vector machines. *Geoderma* **2010**, *154*, 340–347. [[CrossRef](#)]
30. ASTM D2487; Standard Practice for Classification of Soils for Engineering Purposes. ASTM International: West Conshohocken, PA, USA, 2011.
31. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
32. Kumar, D.R.; Samui, P.; Burman, A.; Wipulanusat, W.; Keawsawasvong, S. Liquefaction susceptibility using machine learning based on SPT data. *Intell. Syst. Appl.* **2023**, *20*, 200281. [[CrossRef](#)]
33. Zhao, J.; Li, D.; Jiang, J.; Luo, P. Uniaxial Compressive Strength Prediction for Rock Material in Deep Mine Using Boosting-Based Machine Learning Methods and Optimization Algorithms. *Comput. Model. Eng. Sci.* **2024**, *140*, 275–304. [[CrossRef](#)]
34. Díaz, E.; Spagnoli, G. Gradient boosting trees with Bayesian optimization to predict activity from other geotechnical parameters. *Mar. Georesour. Geotechnol.* **2023**, *42*, 1075–1085. [[CrossRef](#)]
35. Díaz, E.; Salamanca-Medina, E.L.; Tomás, R. Assessment of compressive strength of jet grouting by machine learning. *J. Rock Mech. Geotech. Eng.* **2024**, *16*, 102–111. [[CrossRef](#)]
36. Das, B.; Sobhan, K. A Historical Perspective. In *Principles of Geotechnical Engineering*, 7th ed.; Cengage Learning: Stamford, CT, USA, 2016; pp. 7–35.
37. Das, B.M.; Sobhan, K. *Principles of Geotechnical Engineering*, 7th ed.; Cengage Learning: Stamford, CT, USA, 2017.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.