Jiaxian Li[1], Pengcheng Zhou[1], Junping Ren[1], Yiqing Pu[1], Fanyu Zhang[1], and Chong Wang[1]

[1]College of Civil Engineering and Mechanics, Lanzhou University

March 07, 2024

Abstract

Unfrozen water content (UWC) is a key parameter affecting a variety of soil physical-mechanical properties and processes in frozen soil systems. However, traditional estimation models suffer limitations due to oversimplified assumptions or limited applicable conditions. Given that, there is a compelling need to explore alternative modeling approaches that leverage machine learning (ML) algorithms, which have shown increasing potential in engineering fields. To this end, this study evaluated and compared six widely known ML algorithms (i.e., three ensemble models: RF, LightGBM and XGBoost; and three non-ensemble models: KNN, SVR and BPNN) for modeling UWC based on collected experimental datasets. These algorithms were optimized and evaluated using a framework combining Bayesian optimization and cross-validation to ensure model stability and generalization. The results demonstrated that the ensemble tree-based methods, particularly LightGBM and XGBoost, achieved the highest predictive accuracy and superior overall performance. On the other hand, the nonensemble methods exhibited poorer generalization abilities. Interestingly, during 10-fold cross-validation, consistent underperformance was observed for a particular fold, possibly stemming from the challenges of the data distribution in that fold after random shuffling. The present study highlights the effectiveness of ensemble learning approaches, importance of proper hyperparameter tuning and validation strategies, and intrinsic modeling challenges arising from the difference between the freezing and thawing phase change behaviors. This comprehensive ML model comparison and robust training framework provide valuable guidance on selecting suitable data-driven techniques for modeling frozen soil properties for cold regions hydrogeology and engineering practices.

Unfrozen Water Content Estimation: A Comparison between Ensemble and Non-ensemble Machine Learning Models

Jiaxian Li, Pengcheng Zhou, Junping Ren*, Yiqing Pu, Fanyu Zhang, Chong Wang

College of Civil Engineering and Mechanics, Lanzhou University, Lanzhou 730000, China

*Corresponding author: Junping Ren

Email: renjp@lzu.edu.cn

Abstract:

Unfrozen water content (UWC) is a key parameter affecting a variety of soil physical-mechanical properties and processes in frozen soil systems. However, traditional estimation models suffer limitations due to oversimplified assumptions or limited applicable conditions. Given that, there is a compelling need to explore alternative modeling approaches that leverage machine learning (ML) algorithms, which have shown increasing potential in engineering fields. To this end, this study evaluated and compared six widely known ML algorithms (i.e., three ensemble models: RF, LightGBM and XGBoost; and three non-ensemble models: KNN, SVR and BPNN) for modeling UWC based on collected experimental datasets. These algorithms were optimized and evaluated using a framework combining Bayesian optimization and cross-validation to ensure model stability and generalization. The results demonstrated that the ensemble tree-based methods, particularly LightGBM and XGBoost, achieved the highest predictive accuracy and superior overall performance. On the other hand, the non-ensemble methods exhibited poorer generalization abilities. Interestingly, during 10-fold cross-validation, consistent underperformance was observed for a particular fold, possibly stemming from the challenges of the data distribution in that fold after random shuffling. The present study highlights the effectiveness of ensemble learning approaches, importance of proper hyperparameter tuning and validation strategies, and intrinsic modeling challenges arising from the difference between the freezing and thawing phase change behaviors. This comprehensive ML model comparison and robust training framework provide valuable guidance on selecting suitable data-driven techniques for modeling frozen soil properties for cold regions hydrogeology and engineering practices.

Keywords: *Unfrozen water content; Machine learning; Ensemble learning; Bayesian optimization; Model comparison*

1. Introduction

The freezing of water to form ice is one of the most common phase transformations in the natural environment (Wettlaufer, 1999). At a negative temperature, not all pore water in a soil undergoes transformation into ice; rather, a certain amount of liquid water exists because of capillarity and the surface energy of soil particles, which is termed as unfrozen water (Xu et al., 2001). The relationship between unfrozen water content (UWC) and subzero temperature is typically referred to as the soil-freezing characteristic curve (SFCC) (Ren et al., 2021). The variation of UWC during freezing–thawing process significantly influences the thermal, hydraulic and mechanical properties of frozen soils. It is also often accompanied with water migration (Zhang et al., 2018b), frost heave (Li et al., 2018; Ren et al., 2023a; Pei et al., 2024), and thaw settlement (Zhang and Michalowski, 2015; Liu et al., 2024) of the frozen soil system, which potentially leads to geological disasters as well as poses great threats to the infrastructures and environment in cold regions. Therefore, the accurate determination of UWC in frozen soils is of great scientific and practical importance in cold region hydrogeology and engineering practices.

The UWC in frozen soils depends on plenty of factors, including soil properties (e.g., mineral composition, soil pore size distribution, water content, density, composition and concentration of pore solution), and external conditions which include environmental temperature, pressure, and freezing-thawing and drying-wetting histories (Xu et al., 2001; Tian et al., 2014; Kong et al., 2020). In addition, due to the hysteresis effect between the freezing and thawing branches of SFCC, the UWC at the same subzero temperature often exhibit differences (Zhang et al., 2020; Li JX et al., 2024). The complicated effects of these factors and their intricate interactions on UWC result in difficulties associated with the convenient and precise measurement of UWC in frozen soils, under either the laboratory or in-situ conditions. Therefore, many studies have shifted their focus towards developing UWC estimation models, including empirical relations fitted to experimental data, semi-empirical relations based on soil-water characteristic curve (SWCC), and models derived from various theories (e.g., Mckenzie et al., 2007; Liu and Yu, 2013; Wang C et al., 2017; Bai and Lai, 2018; Li Z et al., 2020).

For example, Anderson and Tice (1972) proposed an empirical model, wherein the UWC is regarded as a simple power function of subzero temperature. However, the model parameters need to be determined by experiments and lack physical meanings (Kong et al., 2020; Wan et al., 2022). To address these limitations, Kong et al. (2020) proposed a piecewise function consisting of a linear equation and a power equation to describe SFCC. In addition, the equation proposed by Anderson and Tice tends to infinity at freezing temperatures close to 0 °C, rendering it unacceptable in numerical modeling of frozen soil behavior. Instead, Michalowski (1993) proposed an exponential equation taking into account the residual UWC, which was adopted by Zhang and Michalowski (2015) for thermo-hydro-mechanical analysis of frost heave and thaw settlement. By combining the simplified Clapeyron equation with the Brooks and Corey (1964) SWCC equation, Sheshukov and Nieber (2011) obtained a relationship for UWC and subzero temperature. Chai et al. (2018) considered the UWC as the sum of unfrozen capillary water and unfrozen bound water, and proposed calculation equations based on the freezing points of these two components. Jin et al. (2020) established a theoretical model for quantify UWC based on independent variables of temperature, specific surface area and electrical double-layer parameters. Wan et al. (2022) employed the premelting theory to investigate the variation in unfrozen water during soil freezing, which provides a new idea to determining UWC. However, these prediction models derived from experimental data and physical theories suffer from a limited application scope, restricting their utility to specific soils and thus falling to meet the requirements for widespread practical application. Addressing these issues is imperative to enhancing our understanding of the complex behavior of water-ice transition in soils during freezing and thawing, potentially paving the way for the development of more precise models for UWC prediction (Zou et al., 2023).

In order to address the aforementioned challenges and develop UWC models with broader applicability, some studies have taken advantage of machine learning (ML) techniques. Shang and Mao (2001) proposed a model based on backpropagation neural network (BPNN) to predict the empirical parameters of the SFCC of Morin Clay under different initial water content, dry density and NaCl concentration. Based on experimental data obtained by nuclear magnetic resonance, Liu et al. (2018) constructed two models using adaptive network fuzzy inference system (ANFIS) and BPNN to predict the UWC of saline soil. Wang Q et al. (2020) proposed a new model to predict the UWC of saline soil based on the combined weighting method and ANFIS. Ren Z et al. (2023) established a model based on the genetic algorithm and BPNN to predict UWC under extremely-low-temperature conditions. Ren et al. (2023b) proposed a BPNN modeling framework for predicting the UWC in various types of soils, based on the collected large amount of experimental data. However, neural networks (NNs) sometimes yield “random” UWC predictions that violate physical mechanisms. To address this issue, Li JX et al. (2024) adopted a constrained monotonic neural network to ensure the predicted UWC decreases as the temperature decreases. However, the algorithms employed in these studies are mainly limited to NNs. In addition, other ML methods, which have been successfully used in the prediction of landslide susceptibility as well as soil properties (e.g., Chen et al., 2017; Baghbani et al., 2022), show potential use in UWC estimation (Nartowska and Sihag, 2024). Therefore, it is imperative to evaluate and compare the performance of various ML algorithms and determine the most suitable algorithmic models for predicting UWC.

In our previous study (Ren et al., 2023b), the hysteresis of SFCC was ignored for simplicity and the freezing and thawing UWC data were combined to train a BPNN model for estimating UWC. A key limitation of this approach stems from the inconsistency between the target values at the same input condition, since the UWC at the same subzero temperature exhibit differences during the freezing and thawing processes, thereby introducing a source of prediction error in the developed model. Therefore, in this study, the experimental UWC data of freezing and thawing branches are separately collected from literature to alleviate this concern. Based on these two datasets, six mainstream ML algorithms (i.e., Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and BPNN) were employed to estimate UWC in frozen soils. The first three algorithms are ensemble learning methods and the rest three are non-ensemble. To ensure model stability and generalization, a framework combing Bayesian optimization and 10-fold cross-validation was used to optimize algorithm hyperparameters and evaluate model performance. The six models were comprehensively compared in terms of their predictive abilities and other quantitative metrics. The advantages and limitations of each approach are critically discussed regarding their suitability for modeling complex soil behavior using freezing and thawing datasets. The results of the present study can guide the selection of suitable data-driven techniques for modeling frozen soil properties. The overall modeling framework is summarized in Fig. 1.

2. Dataset preparation

In the present study, soil physical properties and the UWC data were obtained from the literature. The raw data were extracted from the original plots depicting the UWC-Subzero temperature relations (i.e., SFCC) using the GetData Graph Digitizer. More details regarding the collected data can be found in Ren et al. (2023b). The freezing or thawing process was generally measured in the selected studies, while several studies measured both the freezing and thawing SFCC branches (e.g., Kozlowski and Nartowska, 2013; Ren and Vanapalli, 2019; Teng et al., 2020). However, due to hysteresis between freezing and thawing processes, the same soil sample often exhibits different UWC values at the same subzero temperature. This causes difficulties in ML development since identical inputs corresponding to different outputs in the training data, hindering effective model training and compromising the robustness of the trained model. To avoid this obstacle, the dataset collected from studies that measured both branches was divided into separate freezing and thawing subsets. For studies employing multiple measurement methods, only the data based on NMR measurements were retained, as it is a relatively stable and accurate method to measure UWC without damaging the soil samples (Ren et al., 2020; He et al., 2023). Additionally, for studies measuring UWC under multiple freeze-thaw cycles, only the UWC measurements on either the freezing or thawing branch of the first cycle were included in the database. As a result, two separate datasets were obtained: the freezing branch dataset (FBD) and the thawing branch dataset (TBD). The FBD and TBD comprise 790 and 1410 UWC data points, respectively. All subsequent analysis and discussions in this study will be based on these two separate datasets.

2.1 Data statistical description

Similar to the study by Ren et al. (2023b), the following four factors influencing UWC were considered: specific surface area (SSA), dry density (ρ_d), initial volumetric water content (θ_{ini}), and subzero temperature (*Temp*). The statistical features of the two datasets (i.e., FBD and TBD) are described next. Table 1 summarizes key statistical descriptors of the four input variables and the output (i.e., UWC) for both the freezing and thawing data subsets. The standard deviation, SD, representing the arithmetic square root of the variance, serves as a measure of the extent to which observations deviate from their mean. Skewness, Sk, presents distribution characteristics, with positive Sk suggesting a bias towards larger-than-average data points, while negative Sk signifies a prevalence of observations below the mean. Additionally, Kurtosis (Ku) provides insights into the tail distribution, where high Ku indicates heavy tails and potential outliers, while low kurtosis points to lighter tails with fewer extreme values compared to a normal distribution (Li and Vanapalli, 2022; Li JX et al., 2024). These statistical values are calculated based on Eqs. (1) to (3):

$$SD = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}} \tag{1}$$

$$S_k = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{SD}\right)^3 \tag{2}$$

$$K_u = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{SD}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{3}$$

where n is the total number of a variable, X_i and $\overline{X}$ are the value and mean of the variable, respectively.

Figure 2 presents the histograms as well as kernel density plots of the four variables and the prediction target (i.e., θ_u). The FBD exhibits high variability and non-normal distributions for several key variables. The input feature SSA and the output θ_u exhibit positive skewed distributions, which indicates substantial right-tailed distributions, as evidenced by their Sk values (see Table 1) and kernel density curves. The distribution of θ_{ini} is close to a normal distribution. Meanwhile, ρ_d and $Temp$ show negative skewed distributions with their Sk values below -1, indicating left-tailed shapes. The Ku values of SSA, ρ_d, $Temp$, and θ_u exceeding 3 further demonstrate heavy tails and large values. In comparison, the TBD displays different data distributions. The SSA and θ_u retain strong positive skewness and heavy tails seen in Fig. 2(a) & (d), and the distribution of SSA becomes steeper compared to a normal distribution. The θ_{ini} has a slightly positive skewed distribution while ρ_d shows minor negative skewness. Unlike FBD, the kernel density curves for these two variables in TBD exhibit two peaks (see Fig. 2(b) & (c)), indicating that the distributions of these two variables are more complex or multimodal compared to those in FBD. Although the two datasets share some statistical similarities in their means and SDs, the freezing data overall displays more pronounced non-normal distributions and heavy-tailed characteristics. These statistical differences highlight the unique characteristics inherent between the collected freezing and thawing data.

2.2 Feature importance

The frozen soil is a complicated four-phase system and the amount of unfrozen water in a frozen soil is a regression function of multiple variables. Therefore, it is necessary to identify how much each factor affects the UWC and which factor influences the UWC most. Since UWC depends on the intricate interplay of multiple influencing factors, resulting in a complex non-linear relationship among them, and the Spearman correlation coefficient (SCC) serves as a nonparametric or distribution-free statistical measure to describe the rank of variables (Xiao et al., 2016; Li KQ et al., 2022), herein, Spearman correlation analysis was adopted to analyze the correlation degree between the four input variables and the output. The SCC can be calculated as:

$$SCC = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4}$$

where d_i is the difference between each pair of the ranked variables and n is the total sample size of observations.

Figure 3 shows the correlation relationship among the input variables and the output for the two datasets. Temperature exerts the predominant control on UWC in both the freezing and thawing processes, with SCC values of 0.59 and 0.61, respectively. Dry density (ρ_d) has the least effect on UWC during both processes and was negatively correlated, with SCC values of -0.07 and -0.22, respectively. However, well-defined correlations were not observed between any input variable and the UWC. This is because the interactions between the influencing factors are intricate and not yet fully understood. Therefore, the UWC prediction does not typically depend on any single factor (Li JX et al., 2024). Notably, for the freezing process, θ_{ini} has a greater effect than the SSA, whereas the converse is observed for the thawing process. In the freezing process, the initial liquid water content determines the total amount of water available for phase transition to ice. However, during thawing, the vast majority of the initial water has transformed into the ice phase, therefore its direct influence on UWC diminishes. Instead, the SSA, which quantifies the surface area contact between ice and soil particles, becomes more impactful. A higher SSA provides more surface for conducting heat transfer and water flow during thawing. This shift in the relative importance of influential factors again indicates the differences between the freezing and thawing processes.

3. Models overview and development

3.1 Six machine learning algorithms

A wide variety of ML algorithms have been developed for multivariate regression modeling. For this study, six representative ML algorithms were employed to model and predict UWC: RF, XGBoost, LightGBM, KNN, SVR and BPNN, with the first three being ensemble models and the rest three non-ensemble. The selection of these six ML algorithms is motivated by their diverse strengths and capabilities. The RF, XGBoost, and LightGBM were specifically chosen as ensemble models can achieve stronger predictive performance by combining multiple weak learners. Complementing the ensemble models, we include KNN, SVR, and BPNN, each renowned in ML prediction tasks for their distinct approaches, ensuring a thorough exploration of diverse modeling strategies for UWC prediction. The subsequent sections provide succinct overviews of the underlying principles, as well as general advantages and limitations of each ML algorithms.

3.1.1 Random forest (RF)

The RF method was developed by Breiman (2001) as an expansion of the classification and regression trees technique to provide better performance of prediction results. The RF is an extended algorithm that combines multiple decision trees (DTs) based on the bagging idea of ensemble learning, which enhances basic models' diversity by considering a random set of features at splitting nodes (Li KQ et al., 2022). As schematically illustrated in Fig. 4, the learners (i.e., DTs) are trained separately on the training dataset, and their individual outputs are combined to form the final learning result, with each sample holding equal weight. For regression problems, the final output of RF is the average of the outputs generated by all DTs. The benefits of employing RFs are that the ensembles of trees are used without pruning. In addition, this method is relatively robust to overfitting (Zhang et al., 2020).

3.1.2 Extreme gradient boosting (XGBoost)

The XGBoost is an improved optimization algorithm based on Gradient Boosting Decision Tree (GBDT), as proposed by Chen and Guestrin (2016). In the field of machine learning, it is well recognized that the XGBoost is currently one of the fastest and best open sources boosted tree algorithms. The basic element of XGBoost is the single decision tree and its mechanism is to keep adding and training new trees to fit residuals of last iteration (Dong et al., 2020), as shown in Fig. 5. Compared with GBDT, the XGBoost performs second-order Taylor expansion of the loss function to improve calculation accuracy, and adds a regularization term (i.e., Eq. (6)) to the objective function to prevent overfitting and control the complexity of the model (Yang et al., 2023). Equation 5 evaluates the model “goodness” relative to the original function (Fan et al., 2021).

$$Obj = \sum_{i=1}^{n} l(y_i, \bar{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

where, *Obj* represents the objective function, *l* is the loss function, *K* represents the total number of decision trees, *f_k* represents the complexity of the k^th tree, and *Ω* is the regularization term, which is expressed as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{6}$$

where, *ω* is the score vector, *λ* is the regularization parameter, and *γ* is the mini loss.

3.1.3 Light gradient boosting machine (LightGBM)

Although XGBoost is regarded as a state-of-the-art evaluator with ultra-high performance in both classification and regression tasks, its efficiency and scalability are not satisfactory in the presence of high feature dimensions and large data sizes (Wu, 2020; Wang et al., 2021). In contrast, LightGBM, released by Microsoft in late 2017 (Ke et al., 2017), emerges as a novel gradient boosting technique designed to address the limitations of traditional boosting algorithms, including high memory usage, computational complexity and time consumption (Sun et al., 2022).

Differing from XGBoost, the LightGBM leverages Histogram-based techniques to discretize continuous eigenvalues into multiple integers (also called bins). It performs the gradient accumulation and counting according to the bin where the eigenvalues are located, and then iterates over all the eigenvalues to find the optimal splitting point. This not only improves efficiency but also reduces memory occupation. The discrete split points also have a regularization effect, which could effectively reduce the over-fitting phenomenon for small datasets (Qiu et al., 2023). In addition, based on the Histogram algorithm, the LightGBM implements a leaf-wise algorithm with depth limitation to split the leaf nodes instead of the level-wise technique for growing decision trees. Specifically, in contrast to the level-wise algorithm, which traverses the data once and then splits each leaf of the level, the leaf-wise algorithm first determines which leaf within the level will provide the biggest splitting gain and subsequently performs the split, as shown in Fig. 6. This strategy reduces the complexity of the model, maintains a high-efficiency level, and simultaneously enhances the resistance to overfitting (Hajihosseinlou et al., 2023). Furthermore, the LightGBM uses the gradient-based one-side sampling algorithm and the mutually exclusive feature bundling algorithm to solve the problems of excessive number of samples and features respectively, which further improves the computational efficiency of the model. The LightGBM is also frequently used in data mining competition, such as Kaggle, where it has proven to be a winning solution (Ustuner and Balik, 2019; Cai et al., 2022). For more in-depth explanations of LightGBM, readers can refer to Ke et al. (2017).

3.1.4 K-nearest neighbors (KNN)

The KNN algorithm is one of the simplest ML algorithms in terms of both underlying principles and, often, computational demand. As a nonparametric classifier introduced by Cover and Hart (1967), the KNN is based on labeling the unknown instance using known instances. At the stage of classification for a given new sample, the KNN algorithm searches through all training samples and then computes the distances between the target sample and each training data point to determine the nearest neighbors and produce the classification output (Yamac et al., 2020). Typically, the Euclidean distance algorithm is used to calculate the distances between instances (Araya and Ghezzehei, 2019).

For a simple classification task shown in Fig. 7, when $k = 3$, there are two triangles and one circle in the nearest neighborhood of the unknown class. Consequently, the unknown class is determined to be Class B. While $k = 5$, the category becomes Class A. When applied for regression problems, KNN predicts the value of a new instance by averaging the values of its "k" nearest (i.e., most similar) neighbors in the training data. The KNN is considered as a nonparametric algorithm since it does not assume an underlying data distribution. However, the KNN does not perform any generalization on the training data and retains all data points, which may result in overfitting. Additionally, the need for distance computation of k-nearest neighbors makes the algorithm computationally intensive with large datasets, limiting its scalability (Ray, 2019; Zhao et al., 2022). Moreover, the KNN algorithm is highly sensitive to redundant and irrelevant features and therefore feature selection must be done carefully (Yamac et al., 2020).

3.1.5 Support vector regression (SVR)

Support vector machine (SVM), as a type of generalized linear classifiers proposed by Cortes and Vapnik (1995), is derived from the structural risk minimization hypothesis to minimize both empirical risk and the confidence interval of the learning machine for improved generalization capability. The SVM is developed based on statistical learning theory, the basic idea of which is to map the original datasets from the input space to a high-dimensional or even infinite-dimensional feature space, in order to define a separable hyperplane that maximizes the margin between classes, such that the classification problem becomes simpler in the feature space (Raghavendra and Deka, 2014; Hosseinzadeh et al., 2021). The function that transforms data from input space to feature space is called the kernel function. The SVM model requires the data to be located in this hyperplane as much as possible to minimize the total deviation of all the data from the hyperplane. Additionally, the SVM method uses a small number of support vectors instead of the entire sample space, which makes it easier to calculate the final decision function with improved robustness and efficiency. Compared with complex NNs, the SVM has demonstrated better performance and requires fewer hyperparameters to be tuned while avoiding local minima (Khlosi et al., 2016; Wang F et al., 2020).

Although the SVM was developed to solve classification tasks, it has been extended to regression scenarios (Smola and Schölkopf, 1998), which is known as the support vector regression (SVR). For regression problems, the SVR introduces an ε-insensitive loss function to determine a hyperplane, which allows for some deviation between the predicted and target values without affecting loss calculation. In other words, the loss is calculated only when the absolute value of the difference between predictions and targets is greater than ε (Lu and Wang, 2023). As shown in Fig. 8, values centered on the function and within the error range on either side of it are considered correctly predicted, while only values outside the dash line are incorporated into loss computation and model updating process.

3.1.6 Backpropagation neural network (BPNN)

The BPNN is a well-known learning method for multi-layer feedforward neural network trained by an error backpropagation algorithm (Li J et al., 2012). The BPNN was first proposed by Paul Werbos in 1974 and later popularized by Rumelhart et al. (1986). The BPNN can not only simulate various nonlinear relationships between variables, but also has self-adaptability and self-learning capabilities.

As shown in Fig. 9, a complete BPNN consists of input layer, hidden layers and output layer. The number of neurons in the hidden layers largely affects the performance of BPNNs. Specifically, each neuron in the hidden layer receives the weighted combination of input values from the preceding layer and calculates an output depending on the activation function, which is then propagated as the input to neurons in the next hidden layer. This process can be mathematically represented as:

$$y_j = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{7}$$

where x_i is the value of neurons in the previous layer, w_i represents weights, b represents bias, f is activation function, and y_j is the output of the current neuron.

The training of BPNN includes two key processes: forward propagation of the input signals and backpropagation of error. In the forward propagation, information flows from the input layer to the output layer (Kurt and Kayfeci, 2009). And during backpropagation, the error between the predictions from the forward pass and target values is calculated and then propagated back to the input layer to update weights (Feng et al., 2015). The activation function also plays an important role in the training and performance of the model, as it determines the output of the neurons based on their input. It provides the necessary nonlinearity for the model to represent complex functions. Commonly used activation functions include the rectified linear unit (ReLU) function (Eq. (8)), as used in this study, along with the sigmoid and hyperbolic tangent (tanh) function.

$$\mathrm{ReLU}(x) = max(0, x) \tag{8}$$

Despite its powerful learning ability and popularity, the BPNN has limitations such as being prone to falling into local optimum (Liu et al., 2013), which causes the training of BPNN being more sensitive to the initial network weights (Tongle et al., 2016).

3.2 Bayesian optimization and cross-validation

The above six ML algorithms were employed to construct models based on the freezing branch dataset (FBD) and the thawing branch dataset (TBD). To facilitate subsequent model evaluation, 90 and 160 data points were selected randomly from FBD and TBD respectively, forming the freezing test dataset (FBD_test) and the thawing test dataset (TBD_test). The remaining data were used for models training and validation. Specifically, the rest 700 data points from the original FBD were used as the freezing training-validation dataset (FBD_train-val), while the rest 1250 data points from the original TBD were used as the thawing training-validation dataset (TBD_train-val).

Before models training, all model inputs and outputs should generally be standardized. The purpose of this is to avoid excessive network prediction error due to the large order of magnitude difference between different features (Raju et al., 2020), as well as arguably to make the algorithms converge faster. Therefore, we utilized StandardScaler from the scikit-learn preprocessing library to standardize both the FBD and TBD. The StandardScaler standardizes features by removing the mean and scaling to unit variance. The underlying principle of standardization can be described by the following equation:

$$X_{norm} = \frac{X - \overline{X}}{SD} \tag{9}$$

As noted by Zhang et al. (2020), standardization or normalization is unnecessary for RFs because they are insensitive to the range of inputs. Since DT-based models focus on the distribution of variables and conditional probabilities between them, rather than the raw values, normalization is not required. In fact, for RF, LightGBM, XGBoost and other DT-based models, data normalization has little effect on output results, which has been observed in several studies (Coulston et al., 2012). Therefore, in this study, data standardization was not performed prior to establishing the three DT-based ensemble models.

The predictive performance of ML models depends on the appropriate combinations of hyperparameters, such as the number of regression trees and the number of random variables of nodes (i.e., *Max depth*) in RFs. Hyperparameter optimization is fundamentally a problem of optimizing a specific mapping function over graph-structured configuration space (Zhang et al., 2021). While the significance of hyperparameters is evident, manually exploring the optimum hyperparameter combinations requires experienced insight and can be tedious (Kim et al., 2022). In response to this challenge, Bayesian optimization (BO) emerges as an efficient solution to hyperparameter tuning problem by searching through hyperparameter candidates. The core technique of BO lies in utilizing the prior probability of the objective function and observation points to update the posterior probability distribution and then find the next minimal value point with a more posterior probability distribution and get the optimal hyperparameter through iterations (Zhang et al., 2023). Since new candidates are selected based on the results from previous hyperparameters, the best combination of hyperparameters can be configured in less time and fewer evaluations than grid search or random search (Li and Kanoulas, 2018). Therefore, in this study, the BO was employed to fine-tune the hyperparameters of each model to maximize performance. For the automated search for optimal hyperparameter configurations during model training, we utilized the Hyperopt Python library, leveraging its sequential model-based optimization (SMBO) technique powered by the Tree of Parzen Estimators (TPE) algorithm. This enabled efficient tuning tailored to each model's unique configuration needs.

In addition, we strategically incorporated 10-fold cross-validation (CV) within the BO framework to assess the generalization capabilities of models under each identified hyperparameter combination obtained during the BO process. More specifically, the 10-fold CV process categorizes both FBD_train-val and TBD_train-val into ten equal-sized datasets randomly. In the case of FBD_train-val, a dataset with 630 data points was utilized for training the six ML models, while the remaining 70 data points were for validation. Similarly, for TBD_train-val, a dataset with 1125 and 125 data points was used for training and validation, respectively. This procedure was repeated 10 times with one of the 10 folds served as the validation dataset each time, and 10 validation performance scores were generated for every hyperparameter candidate. The hyperparameter configuration that achieved the highest average score was then selected as the optimal setting for the model. It is undeniable that incorporating 10-fold CV into the framework of BO increases computational cost and runtime for finding the optimal hyperparameters of models. Hence, for the candidate hyperparameters of BPNN, we opted not to set a continuous range but specifying common discrete values. Additionally, the number of Bayesian iterations affects the running times of the whole model (Stephens and Donnelly, 2003). The iteration number was consistently set to 100 in this study to save computational cost.

Three performance indicators were adopted to evaluate the performance of the above six models: coefficient of determination (R^2), root mean square error (RMSE) and mean absolute percentage error (MAPE). The R^2 indicator measures the level of fitness between the target and model prediction values. The RMSE is more sensitive to large errors between the target and prediction, due to the quantification by using squared difference. In contrast, the MAPE demonstrates low sensitivity to outliers, which makes it a suitable indicator for data with anticipated outliers (Huang et al., 2023). The three model indicators can be calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2} \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{\left|\hat{y}_i\right|} \tag{12}$$

where n is the number of samples, y_i is the target value, $\hat{y}_i$ is the prediction value, and $\bar{y}$ is the average value of y.

To assess the stability of the models, we calculated the average R^2, RMSE and MAPE of the 10 folds. The 10-fold average R^2 in validation was used as the score metric for BO. Once the optimal hyperparameter combinations were identified, models with the optimal hyperparameters were saved (to avoid model weights being updated again) and used to predict UWC based on FBD_test and TBD_test. As mentioned previously, these two datasets were separated from the dataset before model training and validation, hence representing new and unseen data for the established models. Since 10-fold CV was incorporated in the BO process, there were 10 test results. The final predicted results were calculated by weighted averaging the test results from each fold, where the weights were the R^2 values on the validation set. This weighted approach was implemented to mitigate potential disparities in performance among different folds, ensuring a balanced representation of the overall model performance. The calculation formula is expressed as:

$$P_{final} = \frac{\sum_{i=1}^{10} P_i \bullet R_i^2}{\sum_{i=1}^{10} R_i^2} \tag{13}$$

where, P_{final} is the final predicted result, P_i is the test result from Fold i, $R^2{}_i$ is the R^2 value on the validation set for Fold i, and i ranges from 1 to 10.

For the purpose of benchmarking the six ML models against the test data, the Taylor diagram was used as an effective tool. It can provide a concise visualization of statistical relationship between the models' predictions and targets, through the correlation coefficient, the centered pattern root-mean-square error (CRMSE) and standard deviation (Taylor, 2001; Hu et al., 2020).

4. Results and analysis

Optimizing hyperparameters is crucial for achieving optimal performance in ML algorithms. However, due to the high dimensionality of the hyperparameter space, exhaustively searching all combinations in the hyperparameter space is computationally expensive and time-consuming. Therefore, only the most influential hyperparameters on model performance were considered and selected for optimization in this study. Table 2 summarizes the key hyperparameters of each ML algorithm and their corresponding optimized values based on FBD_train-val and TBD_train-val. The following analysis utilized these models configured with the optimal hyperparameter combination.

The average R^2 and RMSE values obtained through the 10-fold CV for each model on FBD and TBD during training and validation are summarized in Tables 3 and 4, respectively. The ensemble tree-based models (i.e., RF, LightGBM, XGBoost) demonstrated excellent performance in predicting UWC on both datasets, as evidenced by high R^2 values exceeding 0.97 and low RMSE values below 0.02 during training, indicating strong goodness-of-fit. Among the ensemble models, the LightGBM achieved the highest average training R^2 values of 0.9956 on FBD and 0.9870 on TBD, with RMSE values of 0.0075 and 0.0112, respectively. The XGBoost and RF closely followed with outstanding performance. In contrast, non-ensemble models (KNN, SVR, and BPNN) did not perform as well as the ensemble methods. Their training R^2 scores on FBD ranged from 0.91 to 0.94, with higher RMSE values between 0.025 and 0.035. Specifically, the SVR scored high training R^2 value of 0.9387 and achieved low training RMSE value of 0.0283, outperforming the other two non-ensemble models.

As shown in Fig. 10, the three ensemble methods also performed better than non-ensemble models in validation processes, with R^2 values ranging from 0.85 to 0.90 and low RMSE values. In FBD_val, the LightGBM led with an R^2 value of 0.8601 and an RMSE value of 0.0388. Conversely, for TBD_val, the XGBoost obtained the highest R^2 of 0.8970 with the lowest RMSE. However, non-ensemble models exhibited a decline in performance, with R^2 values dropping to 0.80-0.83 and RMSE values rising to above 0.04, indicating an increase in variance and minor overfitting compared to ensemble models. Among the non-ensemble models, the KNN exhibited the poorest performance on TBD_val (R^2 = 0.7663 and RMSE = 0.0464), despite achieving a training R^2 value of 0.9006. This suggests that KNN may be less effective at capturing the complex relationships between the input variables and UWC, and is more prone to overfitting compared to SVR and BPNN. Summing up the comparison of model performance on training and validation sets, the generalization abilities of ensemble tree-based methods on new data are better than their non-ensemble counterparts.

In order to get more insights on the models' performance, their R^2 results of the 10-fold CV on the two validation datasets (i.e., FBD_val and TBD_val) are depicted in Fig. 11. Interestingly, the validation results show that all models exhibited poorer performance consistently on the 3rd fold compared to others when evaluated on FBD_val (see Fig. 11 (a)). The drop is most pronounced for the KNN, with an R^2 of 0.506 on the 3rd fold compared to its highest R^2 value of 0.9111 – indicating a gap over 0.4. The other models exhibited smaller yet evident decrease on the 3rd fold. In contrast, the worst validation fold generally shifts to the 4th fold when models are evaluated on TBD_val. Notably, the degree of underperformance on the 4th fold improved for all models compared to FBD_val, although the KNN still demonstrated the largest discrepancy on this fold. For the BPNN and XGBoost, the worst validation result on TBD_val occurred in the 10th fold rather than the 4th fold, as shown in Fig. 11(b).

It is worth mentioning that, although the BPNN obtained the lowest training R^2 of 0.9132 and 0.8736 on the two datasets, its R^2 values on the validation sets (i.e., 0.8273 and 0.8161) are the highest among the three non-ensemble models. This implies that BPNN has the lowest degree of overfitting and may have better captured the underlying data patterns. However, it may benefit from additional training data for enhanced performance. In addition, the higher R^2 values of BPNN on the two validation sets compared with that reported in Ren et al. (2023b) (i.e., 0.76) indicates that separately training models based on the freezing and thawing branch datasets could potentially improve model performance.

In Fig. 12, a visual representation compares the target values versus the predictions by the six models on the two test datasets. It is evident that the non-ensemble learning models, particularly KNN, exhibit a larger number of data points deviating from the 1:1 line compared to the ensemble learners, indicating their slightly inferior predictive performance. Figure 13 gives a more straightforward comparison of the six models in terms of their R^2, RMSE and MAPE on the test datasets. Overall, most models achieved satisfactory predictive performance with R^2 values above 0.8 on FBD_test, except for the SVR. Specifically, the LightGBM notably achieved the highest R^2 with the lowest RMSE, outperforming the other five models in terms of accuracy. It also attained the lowest MAPE value of 0.36, indicating a superior ability in minimizing error rates. Followed closely, the XGBoost secured with an R^2 of 0.861, an RMSE of 0.031 and an MAPE of 0.384. However, models' performance on TBD_test exhibited varying results, with the XGBoost achieving the top R^2 and lowest MAPE, clearly demonstrating strongest predictive accuracy on this dataset. The LightGBM also maintained excellent performance with R^2, RMSE and MAPE values of 0.888, 0.031 and 0.306, respectively. Among the other four models, the KNN performed the worst with the lowest R^2, the highest RMSE and the second-highest MAPE. This positions it as the model with the poorest performance and the lowest accuracy on TBD_test. The performance of the SVR and BPNN on this test dataset is similar, as evidenced by their

RMSE values, both of which are 0.041. The significant discrepancy in the model performance between the two test datasets can be attributed to the distinct nature of the datasets' distributions and inherent patterns. Referring to Figs. 12 and 13, the LightGBM performs best in this comparative analysis on both test datasets, outperforming the other five models employed in this study.

Figure 14 is the Taylor diagrams that provide useful diagnostic comparisons between the six ML models' predictions and targets. The reference variable of the Taylor diagrams is the target UWC in test datasets (the REF point on the horizontal axis). It can be seen that the LightGBM obtained the highest correlation coefficient while the RF had the smallest standard deviation on FBD_test. The performance gap between the six models is not large on this test dataset. On TBD_test, however, the models' performance is more discrete, and the KNN achieved the lowest correlation coefficient, far away from the other models as well as the REF point.

5. Discussion

In this study, the collected UWC data were partitioned into separate freezing and thawing datasets. The statistical analysis revealed distinct distributions between the two datasets, suggesting significant differences in their underlying features. This split is justified physically that different mechanisms govern the change of UWC in soil freezing and thawing processes. The former is influenced chiefly by temperature, while the latter becomes more affected by soil particle properties such as SSA. For example, the correlation between SSA and UWC jumps from 0.2 on the FBD to 0.46 on the TBD (see Fig. 3). In addition, although the complicated mechanisms responsible to the hysteresis have not been thoroughly revealed, the difference between the freezing and thawing SFCC branches does manifest, especially in the high subzero temperature range (Tian et al., 2014; Zhou et al., 2018; Ren and Vanapalli, 2020). This means that for the same subzero temperature, the corresponding UWC on the freezing branch is higher than that on the thawing branch. Therefore, the amalgamation of freezing-thawing UWC dataset for training ML models introduces potential risks of ambiguous input-output mappings. That is, the same input corresponds different targets in the training dataset, which may compromise the stability and robustness of models during training. Additionally, this uncertainty could lead to notable fluctuations in the models' prediction for a given input, or, the trained model may struggle to generalize to new, unseen data, as it is hard for it to produce correct output for the same input scenarios. This limitation can adversely affect the model's performance when applied to real-world situations. Hence, the reasonable spilt of the freezing and thawing data enables the ML models to better capture the inherent laws of changes in UWC during the freezing and thawing processes, and improve the accuracy of prediction.

The 10-fold CV was integrated within BO to determine the optimal hyperparameter configuration for each model. This framework enables maximizing the potential of each optimized model for fair comparison rather than relying solely on the performance based on a single random split. In other words, the cross-validation is effective in avoiding the impacts of the randomness of dataset division and ensuring the robustness of the trained model. And to a certain extent, it can also help mitigate overfitting and underfitting. However, the application of the 10-fold CV in BO does impose additional computational expense which may become prohibitive for inherently slower models like NNs. Therefore, when optimizing the hyperparameters of BPNN in this study, we did not set a continuous parameter range like the other five models. Instead, we opted for specific, predetermined values, such as restricting the number of hidden layers to discrete options like 1, 2, and 3, depending on the question investigated. This strategy was made to strike a good balance between the efficient exploration of the hyperparameter space and the reduction of generalization error for subsequent comparison between models.

The noteworthy underperformance on the third fold during the 10-fold CV of models built on FBD warrants further investigation. One potential explanation is that in this fold, the validation data distribution pattern is rather distinctive, encompassing a greater number of particular samples, such as more outliers or noisy data, compared to the training data. That is, models may be less capable of generalizing to the validation data in Fold 3 due to insufficient similar samples in the training data of this fold. In contrast, during the CV of models based on TBD, most models exhibited poorest performance on the fourth fold. This shift partially highlights the difference in data distribution between the freezing and thawing datasets. The above phenomenon could also be attributed to some problems in parameters and hyperparameters tuning for the models. It is plausible that models require different parameter settings for optimal performance based on underlying distribution and pattern of the training data in Fold 3, which emphasizes the importance of carefully selecting and tuning parameters during ML tasks. Furthermore, this observation reflects the potential for models to experience notable performance drops under certain dataset spilt, which demonstrates the necessity of using the 10-fold CV in this study to assess model robustness and generalizability.

Moreover, it is also possible that subtle overfitting effects may have occurred which negatively impact models' performance during the validation process. Taking the performance of models based on FBD_train-val as an example, the high R^2 and low RMSE values during the training process on the Fold 3 imply adequate model fitting. As shown in Table 5, however, the performance of all models on unseen data in the validation set has significantly decreased, indicating a potential occurrence of overfitting. This overfitting tendency is not exclusive to the validation set, but extending to the test sets (i.e., FBD_test and TBD_test). Comparing Table 3 and Figure 13 reveals a notable decline in predictive performance of models on the two test sets compared to the training sets, accompanied by a significant increase in errors, where the R^2 difference is approximately 10%. Despite employing the weighted approach to obtain the final test results, this potential overfitting phenomenon still occurs in two test sets, which would be mitigated through the augmentation of dataset size by collecting additional data or the introduction of regularization techniques to the models.

Among the six ML models, the KNN exhibited its poorest performance on Fold 3 of the FBD_val and Fold 4 of the TBD_val, displaying notable disparities compared to other folds. This discrepancy suggests that the performance of KNN may be significantly influenced by the local structure of the data. As a nonparametric algorithm, KNN relies solely on a few nearest training samples (i.e., its "neighbors"), making it susceptible to the influence of outliers when the value of *k* is small (Abu Alfeilat et al., 2019). If a particular fold contains data with substantial variations or non-uniform distribution in specific regions, KNN may exhibit poor performance in that fold. Furthermore, the performance of KNN may be highly influenced by the selection of hyperparameters, such as the number of neighbors (i.e., *k*). Different folds may necessitate varying numbers of neighbors to adapt to changes in the local data structure. Consequently, considerable variations in performance may be observed in the 10-fold CV. Therefore, when employing the KNN algorithm, it is suggested that special attention should be paid to the handling of outliers and noise.

The LightGBM and XGBoost, as highly optimized gradient boosting algorithms, demonstrated superior predictive power on both FBD and TBD. These algorithms iteratively fit new models to emphasize previously mispredicted instances, thereby incrementally optimizing the ensemble as a whole. Although the overall performance of the RF is not as good as the above two boosting models, its performance is superior to that of the non-ensemble models. This is expected as ensemble learning algorithms like random forests and boosting methods (e.g., LightGBM and XGBoost) combine multiple weaker models to create an overall stronger model, reducing variance (Skurichina and Duin, 2002; Ferreira and Figueiredo, 2012). Specifically, taking the RF as an example, it averages predictions from an ensemble of decorrelated decision trees grown on random subsets of the data and features, which helps reduce variance relative to a single decision tree model. In contrast, the non-ensemble methods such as KNN, SVR, and BPNN did not exhibit such predictive advantages. The predictions of KNN rely on the average of the *k* nearest neighbors. However, in the study, Bayesian optimization results indicate that the optimal values for the number of neighbors (i.e., *k* value) in KNN are 2 and 3 on FBD_train and TBD_train, respectively (refer to Table 2). Such relatively small *k* values may make models more susceptible to noisy data and outliers, so that KNN yielded unreasonable predictions. The SVR and BPNN, while possessing universal approximation properties, are prone to overfitting given challenges associated with hyperparameter optimization and lack of ensemble effect. However, as mentioned in Section 4, the BPNN did display competitive capability in UWC prediction among the three non-ensemble algorithms. This can be attributed to its robust and strong power to model complex nonlinear relationships, stemming from its multilayer structure and the application of the backpropagation algorithm during training.

In summary, the ensemble approaches provided the most effective and robust solutions for the prediction task in this study because of their ability to synergistically combine multiple simple basic learners, especially the gradient boosting methods. The three non-ensemble models manifested relatively poorer performance, even though the 10-fold CV strategy ensured their robustness and stability. Further hyperparameter tuning and diverse ensemble techniques could help the non-ensemble models boost predictive accuracy and achieve better generalization, on the assignment of estimating UWC in frozen soils. Future work may explore how to effectively connect the freezing and thawing sub-models into a unified framework to capture the complexity of soil behaviors.

6. Summary

In this study, the UWC data collected from the literature was partitioned into separate freezing and thawing datasets. Based on the two datasets, six machine learning models were developed and evaluated for estimating UWC in frozen soils, including RF, LightGBM, XGBoost, KNN, SVR and BPNN. To ensure the robustness and generalizability of models, the integrated 10-fold CV and BO framework was employed to assess the stability of models and identify optimal hyperparameters across different data splits.

The results demonstrated that the three ensemble models (RF, LightGBM and XGBoost) achieved superior accuracy and satisfactory generalization abilities, owing to their synergistic integration of multiple basic learners. The LightGBM and XGBoost displayed the top prediction power on both the freezing and thawing test datasets. Despite slightly lower scores, the RF also exhibited reliable performance. On the other hand, the non-ensemble algorithms including KNN, SVR and BPNN performed relatively poorer in predictive accuracy compared to ensemble models, as evidenced by their lower R^2 and larger RMSE during both training and validation. Among the non-ensembles, the BPNN showcased relatively robust modeling proficiency, which attributes to its nonlinear approximation strengths. Overall, the non-ensemble methods lagged behind their ensemble counterparts.

Findings highlight the superiority and effectiveness of ensemble learning approaches, especially gradient boosting trees, for the UWC estimation task in the study. The present results provide useful guidance on selecting and applying advanced machine learning techniques for modeling frozen soil properties and behaviors during different processes. It underscores the importance of proper validation strategies and accounting for distinct freezing/thawing phase change behaviors when developing data-driven models for cold regions hydrogeology and engineering practices.

Acknowledgments

The authors gratefully acknowledge the financial support from the National Key R&D Program of China (2022YFC3003401), the National Natural Science Foundation of China (42301141), the Natural Science Foundation of Gansu (22JR5RA447), the Fundamental Research Funds for the Central Universities (lzujbky-2021-sp67), and the Open Fund from the CAS State Key Laboratory of Frozen Soil Engineering (SKLFSE202102).

Competing Interests

The authors declare there are no competing interests.

Data Availability Statement

The data that support the findings of this study are available upon reasonable request.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big data, 7(4), 221-248.

Anderson, D. M., & Tice, A. R. (1972). Predicting unfrozen water contents in frozen soils from surface area measurements. Highway Research Record, 393(2), 12-18.

Araya, S. N., & Ghezzehei, T. A. (2019). Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Resources Research, 55(7), 5715-5737.

Baghbani, A., Choudhury, T., Costa, S., & Reiner, J. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. Earth-Science Reviews, 228, 103991.

Bai, R., Lai, Y., Zhang, M., & Yu, F. (2018). Theory and application of a novel soil freezing characteristic curve. Applied Thermal Engineering, 129, 1106-1114.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

Brooks, R. H., & Corey, A. T. (1966). Properties of porous media affecting fluid flow. Journal of The Irrigation and Drainage Division, 92(2), 61-88.

Cai, W., Wei, R., Xu, L., & Ding, X. (2022). A method for modelling greenhouse temperature using gradient boost decision tree. Information Processing in Agriculture, 9(3), 343-354.

Chai, M., Zhang, J., Zhang, H., Mu, Y., Sun, G., & Yin, Z. (2018). A method for calculating unfrozen water content of silty clay with consideration of freezing point. Applied Clay Science, 161, 474-481.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., ... & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena, 151, 147-160.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297.

Coulston, J. W., Moisen, G. G., Wilson, B. T., Finco, M. V., Cohen, W. B., & Brewer, C. K. (2012). Modeling percent tree canopy cover: a pilot study. Photogrammetric Engineering and Remote Sensing, 78(7), 715-727.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

Dong, W., Huang, Y., Lehane, B., & Ma, G. (2020). XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Automation in Construction, 114, 103155.

Fan, J., Zheng, J., Wu, L., & Zhang, F. (2021). Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models. Agricultural Water Management, 245, 106547.

Feng, Q., Zhang, J., Zhang, X., & Wen, S. (2015). Proximate analysis based prediction of gross calorific value of coals: A comparison of support vector machine, alternating conditional expectation and artificial neural network. Fuel Processing Technology, 129, 120-129.

Ferreira, A. J., & Figueiredo, M. A. (2012). Boosting algorithms: A review of methods, theory, and applications. Ensemble Machine Learning: Methods and applications, 35-85.

Hajihosseinlou, M., Maghsoudi, A., & Ghezelbash, R. (2023). A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. Natural Resources Research, 1-22.

Hayashi, M. (2013). The cold vadose zone: Hydrological and ecological significance of frozen-soil processes. Vadose Zone Journal, 12(4).

He, Y., Xu, Y., Lv, Y., Nie, L., Kong, F., Yang, S., ... & Li, T. (2023). Characterization of unfrozen water in highly organic turfy soil during freeze–thaw by nuclear magnetic resonance. Engineering Geology, 312, 106937.

Hosseinzadeh, A., Moeinaddini, A., & Ghasemzadeh, A. (2021). Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model. Journal of Safety Research, 77, 151-160.

Hu, G., Zhao, L., Zhu, X., Wu, X., Wu, T., Li, R., ... & Hao, J. (2020). Review of algorithms and parameterizations to determine unfrozen water content in frozen soil. Geoderma, 368, 114277.

Huang, Y., Wang, Y., Wang, P., & Lai, Y. (2023). An XGBOOST predictive model of void ratio in sandy soils with shear-wave velocity as major input. Transportation Geotechnics, 42, 101100.

Jin, X., Yang, W., Gao, X., Zhao, J. Q., Li, Z., & Jiang, J. (2020). Modeling the unfrozen water content of frozen soil based on the absorption effects of clay surfaces. Water Resources Research, 56(12), e2020WR027482.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., & Cornelis, W. M. (2016). Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil.

European Journal of Soil Science, 67(3), 276-284.

Kim, D., Kwon, K., Pham, K., Oh, J. Y., & Choi, H. (2022). Surface settlement prediction for urban tunneling using machine learning algorithms with Bayesian optimization. Automation in Construction, 140, 104331.

Kong, L., Wang, Y., Sun, W., & Qi, J. (2020). Influence of plasticity on unfrozen water content of frozen soils as determined by nuclear magnetic resonance. Cold Regions Science and Technology, 172, 102993.

Kozlowski, T., & Nartowska, E. (2013). Unfrozen water content in representative bentonites of different origin subjected to cyclic freezing and thawing. Vadose Zone Journal, 12(1), vzj2012-0057.

Kurt, H., & Kayfeci, M. (2009). Prediction of thermal conductivity of ethylene glycol–water solutions by using artificial neural networks. Applied Energy, 86(10), 2244-2248.

Li, D., & Kanoulas, E. (2018, February). Bayesian optimization for optimizing retrieval systems. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 360-368).

Li, J., Cheng, J. H., Shi, J. Y., & Huang, F. (2012). Brief introduction of back propagation (BP) neural network algorithm and its improvement. In Advances in Computer Science and Information Engineering: Volume 2 (pp. 553-558). Springer Berlin Heidelberg.

Li, J., Ren, J., Fan, X., Zhou, P., Pu, Y., & Zhang, F. (2024). Estimation of unfrozen water content in frozen soils based on data interpolation and constrained monotonic neural network. Cold Regions Science and Technology, 218, 104094.

Li, K. Q., Liu, Y., & Kang, Q. (2022). Estimating the thermal conductivity of soils using six machine learning algorithms. International Communications in Heat and Mass Transfer, 136, 106139.

Li, S., Zhang, M., Pei, W., & Lai, Y. (2018). Experimental and numerical simulations on heat-water-mechanics interaction mechanism in a freezing soil. Applied Thermal Engineering, 132, 209-220.

Li, Y., & Vanapalli, S. K. (2022). Prediction of Soil–Water Characteristic Curves of Fine-grained Soils Aided by Artificial Intelligent Models. Indian Geotechnical Journal, 52(5), 1116-1128.

Li, Z., Chen, J., & Sugimoto, M. (2020). Pulsed NMR measurements of unfrozen water content in partially frozen soil. Journal of Cold Regions Engineering, 34(3), 04020013.

Liu, M., Wang, M., Wang, J., & Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical, 177, 970-980.

Liu, Q., Cai, G., Zhou, C., Yang, R., & Li, J. (2024). Thermo-hydro-mechanical coupled model of unsaturated frozen soil considering frost heave and thaw settlement. Cold Regions Science and Technology, 217, 104026.

Liu, Y., Wang, Q., Zhang, X., Song, S., Niu, C., & Shangguan, Y. (2018). Using ANFIS and BPNN methods to predict the unfrozen water content of saline soil in Western Jilin, China. Symmetry, 11(1), 16.

Liu, Z., & Yu, X. (2013). Physically based equation for phase composition curve of frozen soils. Transportation Research Record, 2349(1), 93-99.

Lu, Y., & Wang, G. (2023). A load forecasting model based on support vector regression with whale optimization algorithm. Multimedia Tools and Applications, 82(7), 9939-9959.

Michalowski, R. L. (1993). A constitutive model of saturated soils for frost heave simulations. Cold Regions Science and Technology, 22(1), 47-63.

McKenzie, J. M., Voss, C. I., & Siegel, D. I. (2007). Groundwater flow with energy transport and water–ice phase change: numerical simulations, benchmarks, and application to freezing in peat bogs. Advances in Water Resources, 30(4), 966-983.

Nartowska, E., & Sihag, P. (2024). Soft Computing Techniques for Predicting Unfrozen Water Content in Copper Contaminated Bentonites: A Comparative Study of Gaussian Process Regression, Support Vector Machine, and Random Forest Models. DOI: 10.2139/ssrn.4706976

Pei, Q. Y., Zou, W. L., Han, Z., Wang, X. Q., & Xia, X. L. (2024). Compression behaviors of a freeze–thaw impacted clay under saturated and unsaturated conditions. Acta Geotechnica, 1-18.

Qiu, Y., Wang, J., & Li, Z. (2023). Personalized HRTF prediction based on LightGBM using anthropometric data. China Communications.

Raghavendra, S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. Applied Soft Computing, 19, 372-386.

Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August). Study the influence of normalization/transformation process on the accuracy of supervised classification. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp.

729-735). IEEE.

Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

Ren, J., Fan, X., Yu, X., Vanapalli, S., & Zhang, S. (2023b). Use of an artificial neural network model for estimation of unfrozen water content in frozen soils. Canadian Geotechnical Journal, 60(8), 1234-1248.

Ren, J., Vanapalli, S. K., & Han, Z. (2017). Soil freezing process and different expressions for the soil-freezing characteristic curve. Sciences in Cold and Arid Regions, 9(3), 221-228.

Ren, J., Zhang, S., Ishikawa, T., Li, S., & Wang, C. (2023a). The frost heave characteristics of a coarse-grained volcanic soil quantified by particle image velocimetry. Geoderma, 430, 116352.

Ren, J., Zhang, S., Wang, C., Ishikawa, T., & Vanapalli, S. K. (2021). The Measurement of Unfrozen Water Content and SFCC of a Coarse-Grained Volcanic Soil. Journal of Testing and Evaluation, 51(1).

Ren, Z., Liu, J., Jiang, H., & Wang, E. (2023). Experimental study and simulation for unfrozen water and compressive strength of frozen soil based on artificial freezing technology. Cold Regions Science and Technology, 205, 103711.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

Shang, S., & Mao, X. (2001). Prediction Model of Soil Freezing Temperature and Unfrozen Water Content Based on Back-Propagation Neural Network. Journal of Glaciology and Geocryology, 23(4): 414-418. (in Chinese)

Sheshukov, A. Y., & Nieber, J. L. (2011). One-dimensional freezing of nonheaving unsaturated soils: Model formulation and similarity solution. Water Resources Research, 47(11).

Skurichina, M., & Duin, R. P. (2002). Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications, 5, 121-135.

Smola, A. J., & Schölkopf, B. (1998). Learning with kernels (Vol. 4). Berlin, Germany: GMD-Forschungszentrum Informationstechnik.

Stephens, M., & Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American Journal of Human Genetics, 73(5), 1162-1169.

Sun, Y., Zhang, F., Lin, H., & Xu, S. (2022). A Forest Fire Susceptibility Modeling Approach Based on Light Gradient Boosting Machine Algorithm. Remote Sensing, 14(17), 4362.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106(D7), 7183-7192.

Teng, J., Kou, J., Yan, X., Zhang, S., & Sheng, D. (2020). Parameterization of soil freezing characteristic curve for unsaturated soils. Cold Regions Science and Technology, 170, 102928.

Tian, H., Wei, C., Wei, H., & Zhou, J. (2014). Freezing and thawing characteristics of frozen soils: Bound water content and hysteresis phenomenon. Cold Regions Science and Technology, 103, 74-81.

Tongle, X., Yingbo, W., & Kang, C. (2016). Tailings saturation line prediction based on genetic algorithm and BP neural network. Journal of Intelligent & Fuzzy Systems, 30(4), 1947-1955.

Ustuner, M., & Balik Sanli, F. (2019). Polarimetric target decompositions and light gradient boosting machine for crop classification: A comparative evaluation. ISPRS International Journal of Geo-Information, 8(2), 97.

Wan, X., Pei, W., Lu, J., Zhang, X., Yan, Z., & Pirhadi, N. (2022). Prediction of the unfrozen water content in soils based on premelting theory. Journal of Hydrology, 608, 127505.

Wang, C., Lai, Y., & Zhang, M. (2017). Estimating soil freezing characteristic curve based on pore size distribution. Applied Thermal Engineering, 124, 1049-1060.

Wang, F., Shi, Z., Biswas, A., Yang, S., & Ding, J. (2020). Multi-algorithm comparison for predicting soil salinity. Geoderma, 365, 114211.

Wang, L., Wu, J., Zhang, W., Wang, L., & Cui, W. (2021). Efficient seismic stability analysis of embankment slopes subjected to water level changes using gradient boosting algorithms. Frontiers in Earth Science, 9, 807317.

Wang, Q., Liu, Y., Zhang, X., Fu, H., Lin, S., Song, S., & Niu, C. (2020). Study on an AHP-entropy-ANFIS model for the prediction of the unfrozen water content of sodium-bicarbonate-type salinization frozen soil. Mathematics, 8(8), 1209.

Wettlaufer, J. S. (1999). Ice surfaces: macroscopic effects of microscopic structure. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 357(1763), 3403-3425.

Wu, C. Z. (2020). Soil undrained shear strength Prediction Based on XGBoost and LightGBM Model.

Chongqing University. (in Chinese)

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. Concurrency and Computation: Practice and Experience, 28(14), 3866-3878.

Yamac, S. S., Şeker, C., & Negiş, H. (2020). Evaluation of machine learning methods to predict soil moisture constants with different combinations of soil input data for calcareous soils in a semiarid area. Agricultural Water Management, 234, 106121.

Yang, Z., He, Q., Miao, S., Wei, F., & Yu, M. (2023). Surface Soil Moisture Retrieval of China Using Multi-Source Data and Ensemble Learning. Remote Sensing, 15(11), 2786.

Zhang, M., Shi, W., & Xu, Z. (2020). Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data. Hydrology and Earth System Sciences, 24(5), 2505-2526.

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geoscience Frontiers, 12(1), 469-477.

Zhang, X., Dai, C., Li, W., & Chen, Y. (2023). Prediction of compressive strength of recycled aggregate concrete using machine learning and Bayesian optimization methods. Frontiers in Earth Science, 11, 1112105.

Zhang, X., Zhang, M., Pei, W., & Lu, J. (2018b). Experimental study of the hydro-thermal characteristics and frost heave behavior of a saturated silt within a closed freezing system. Applied Thermal Engineering, 129, 1447-1454.

Zhang, Y., & Michalowski, R. L. (2015). Thermal-hydro-mechanical analysis of frost heave and thaw settlement. Journal of Geotechnical and Geoenvironmental Engineering, 141(7), 04015027.

Zhao, T., Liu, S., Xu, J., He, H., Wang, D., Horton, R., & Liu, G. (2022). Comparative analysis of seven machine learning algorithms and five empirical models to estimate soil thermal conductivity. Agricultural and Forest Meteorology, 323, 109080.

Zhou, Z., Ma, W., Zhang, S., Mu, Y., & Li, G. (2018). Effect of freeze-thaw cycles in mechanical behaviors of frozen loess. Cold Regions Science and Technology, 146, 9-18.

Zou, Y., Jiang, H., Wang, E., Liu, X., & Du, S. (2023). Variation and prediction of unfrozen water content in different soils at extremely low temperature conditions. Journal of Hydrology, 624,

129900.

Table 1. Statistical description of the two datasets

Dataset	Variable	Unit	Mean	SD	Sk	Ku
FBD	SSA	m^2/g	90.35	111.433	3.421	13.872
	θ_{ini}	m^3/m^3	0.35	0.143	0.721	1.148
	ρ_d	g/cm^3	1.54	0.304	-2.317	6.265
	$Temp$	°C	-7.78	6.519	-1.583	6.437
	θ_u	m^3/m^3	0.13	0.112	1.723	4.510
TBD	SSA	m^2/g	63.12	103.229	4.585	23.767
	θ_{ini}	m^3/m^3	0.38	0.170	0.172	-0.386
	ρ_d	g/cm^3	1.38	0.336	-0.840	0.465
	$Temp$	°C	-4.93	5.199	-1.511	1.948
	θ_u	m^3/m^3	0.11	0.098	1.665	3.341

Table 2. Key hyperparameters for the six ML models

Model	Key hyperparameters	Optimal values	
		Freezing	Thawing
RF	n_estimators	664	789
	max_depth	12	20
	min_samples_split	2	2
	min_samples_leaf	1	1
	max_features	1	1
LightGBM	n_estimators	768	513
	max_depth	7	7
	num_leaves	70	30
	min_child_samples	2	13
	subsample	0.66	0.73
	reg_alpha	0.02	0.02
	reg_lambda	10.17	1.14
XGBoost	n_estimators	200	726
	max_depth	7	7
	learning_rate	0.52	0.14
	subsample	0.84	0.26
	reg_alpha	0.06	0.02
	reg_lambda	27.19	17.78
KNN	algorithm	2	2
	n_neighbors	2	3
SVR	kernel function	RBF	RBF
	c	6.9	7.2
	epsilon	0.007	0.049
	gamma	6.03	7.00
BPNN	n_layer	1	1
	n_hid	64	32
	lr	0.05	0.02
	batch_size	128	128
	activation function	ReLu	ReLu
	epochs	100	100

Table 3. Models' performance on FBD

Process / Models	Training		Validation	
	R^2	RMSE	R^2	RMSE
RF	0.9743	0.0183	0.8551	0.0413
LightGBM	**0.9956**	**0.0075**	**0.8601**	**0.0388**
XGBoost	0.9928	0.0097	0.8526	0.0406
KNN	0.9352	0.0291	0.8091	0.0460
SVR	0.9387	0.0283	0.8209	0.0459
BPNN	0.9132	0.0336	0.8273	0.0451

Table 4. Models' performance on TBD

Process / Models	Training		Validation	
	R^2	RMSE	R^2	RMSE
RF	0.9833	0.0127	0.8766	0.0338
LightGBM	**0.9870**	**0.0112**	0.8806	0.0331
XGBoost	0.9869	0.0113	**0.8970**	**0.0311**
KNN	0.9006	0.0311	0.7663	0.0464
SVR	0.9276	0.0266	0.8140	0.0415
BPNN	0.8736	0.0351	0.8161	0.0415

Table 5. Models' performance on Fold 3 based on FBD

Process / Models	Training		Validation	
	R^2	RMSE	R^2	RMSE
RF	0.9212	0.0330	0.6501	0.0458
LightGBM	0.9964	0.0070	0.5570	0.0515
XGBoost	0.9937	0.0093	0.6442	0.0462
KNN	0.9400	0.0288	0.5060	0.0544
SVR	0.9428	0.0282	0.7058	0.0420
BPNN	0.9212	0.0330	0.6501	0.0458

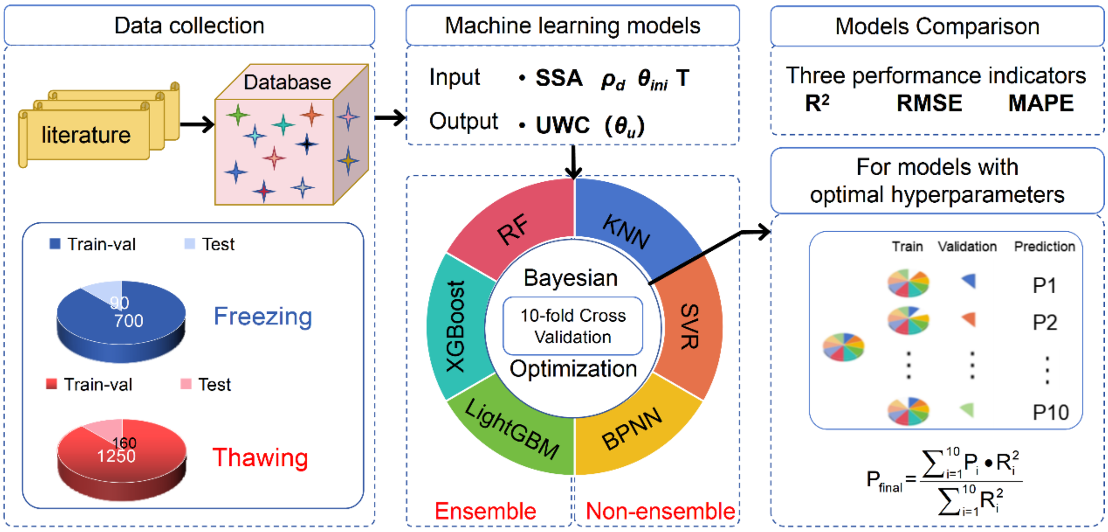


Fig. 1 Framework for unfrozen water content estimation

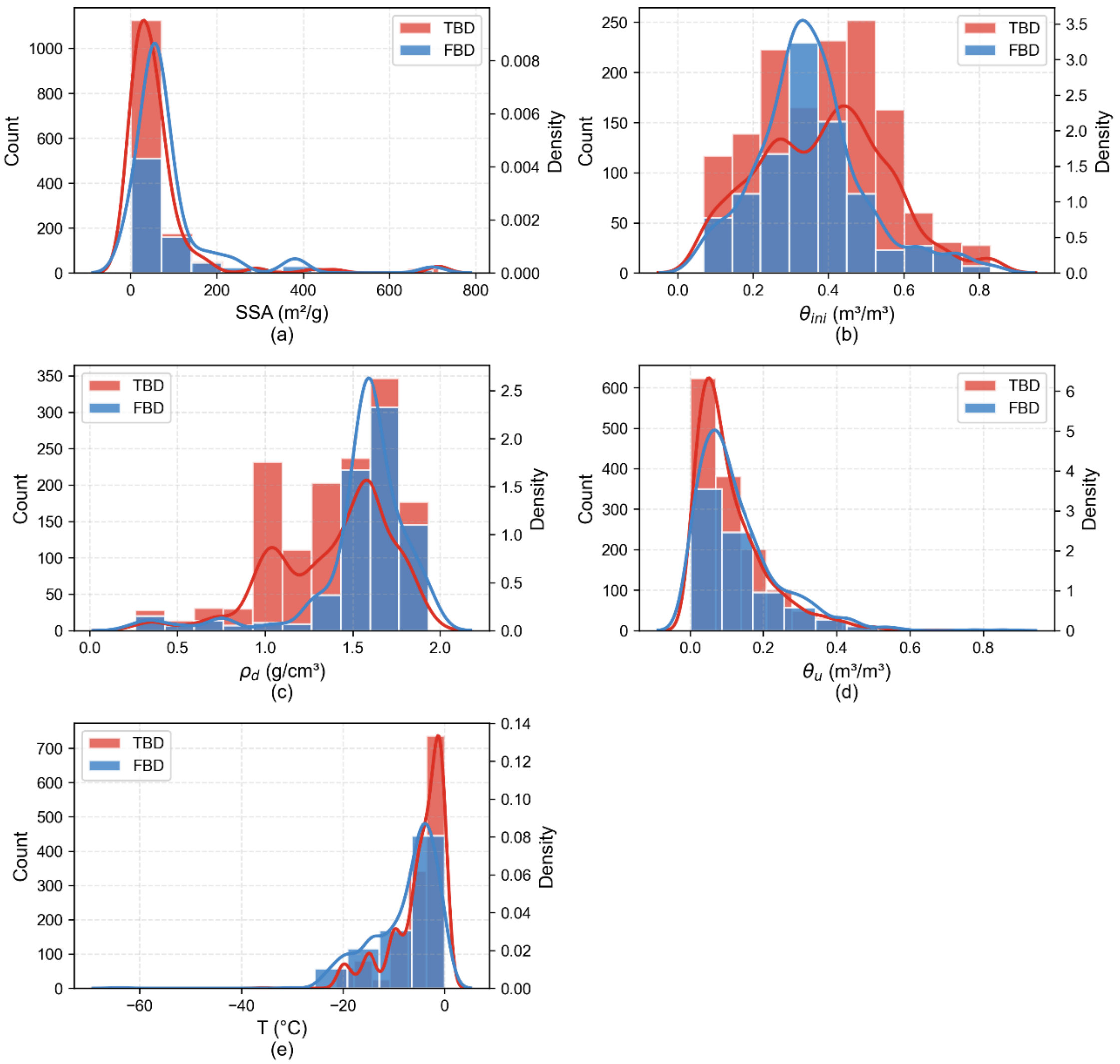


Fig. 2 Histogram plots of the input variables and output

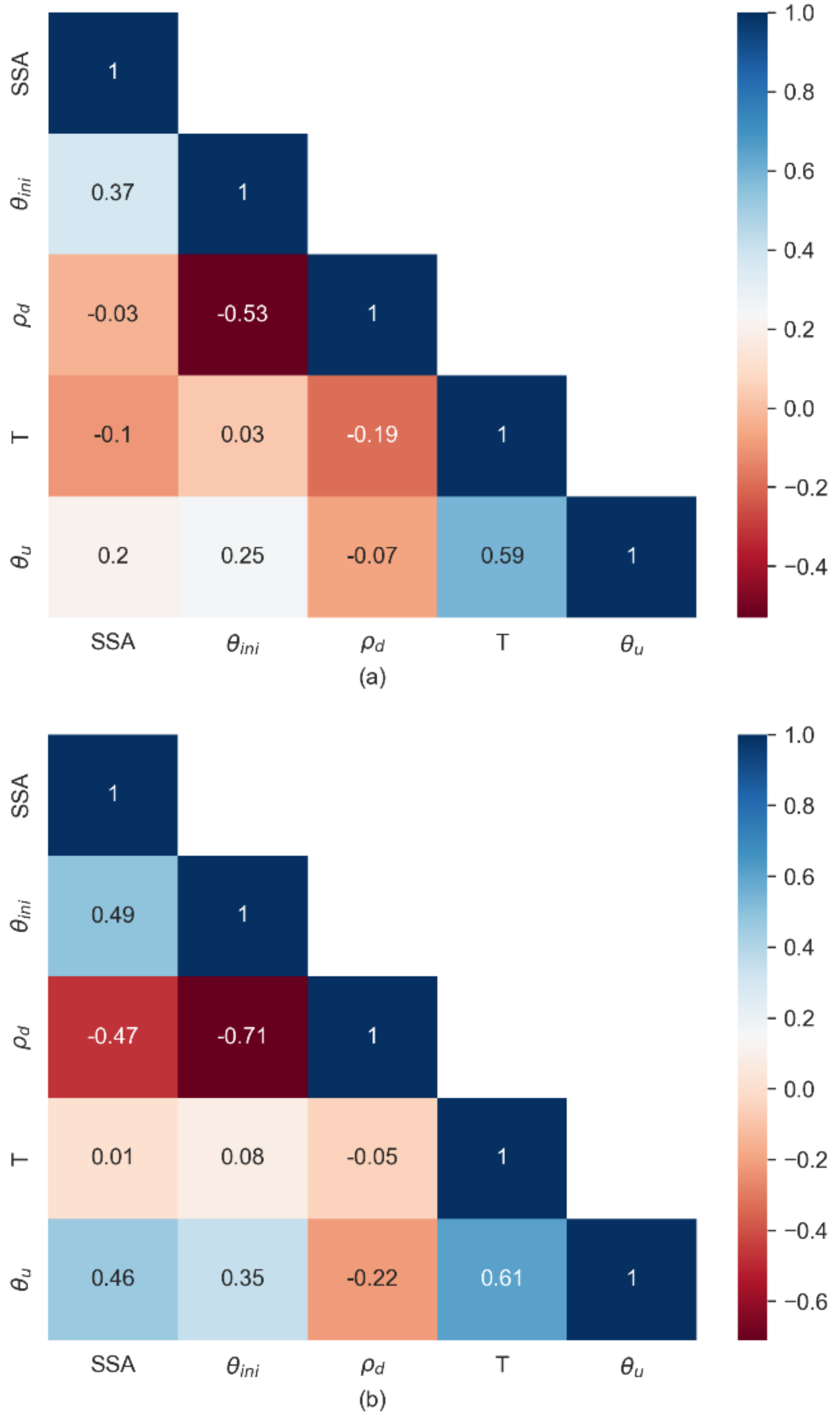


Fig. 3 Spearman correlation coefficient heat map among variables on (a) FBD and (b) TBD

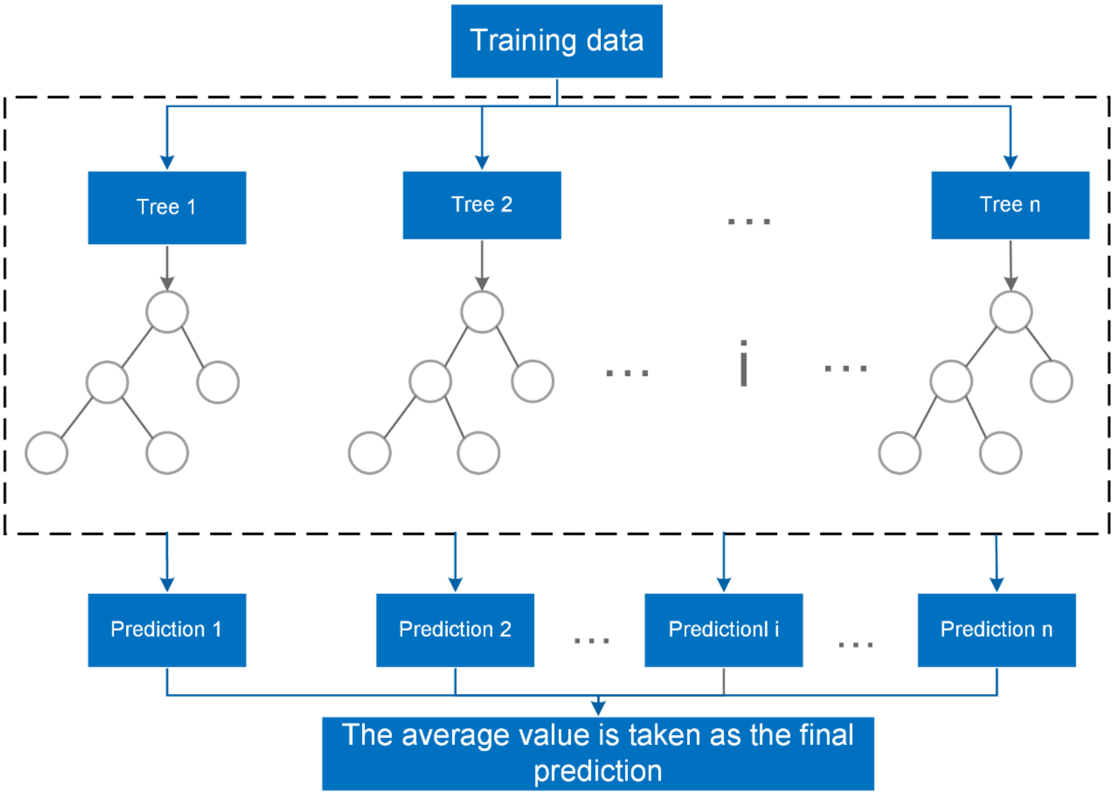


Fig. 4 The structure of random forest

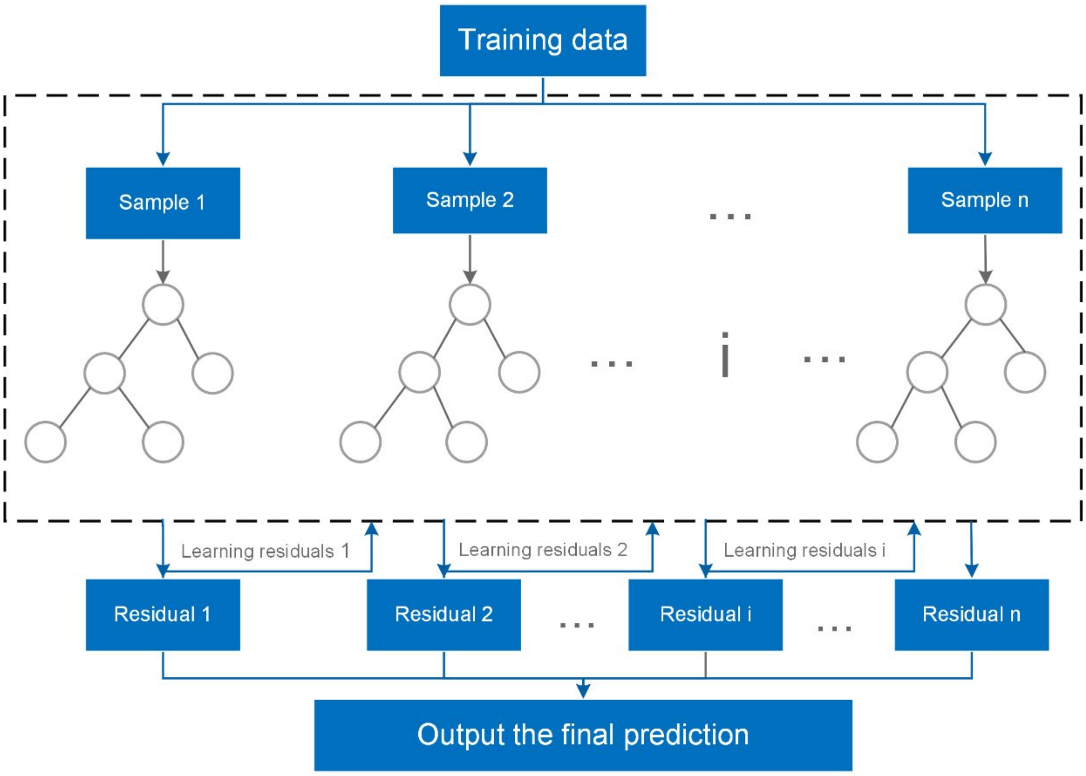


Fig. 5 The structure of XGBoost

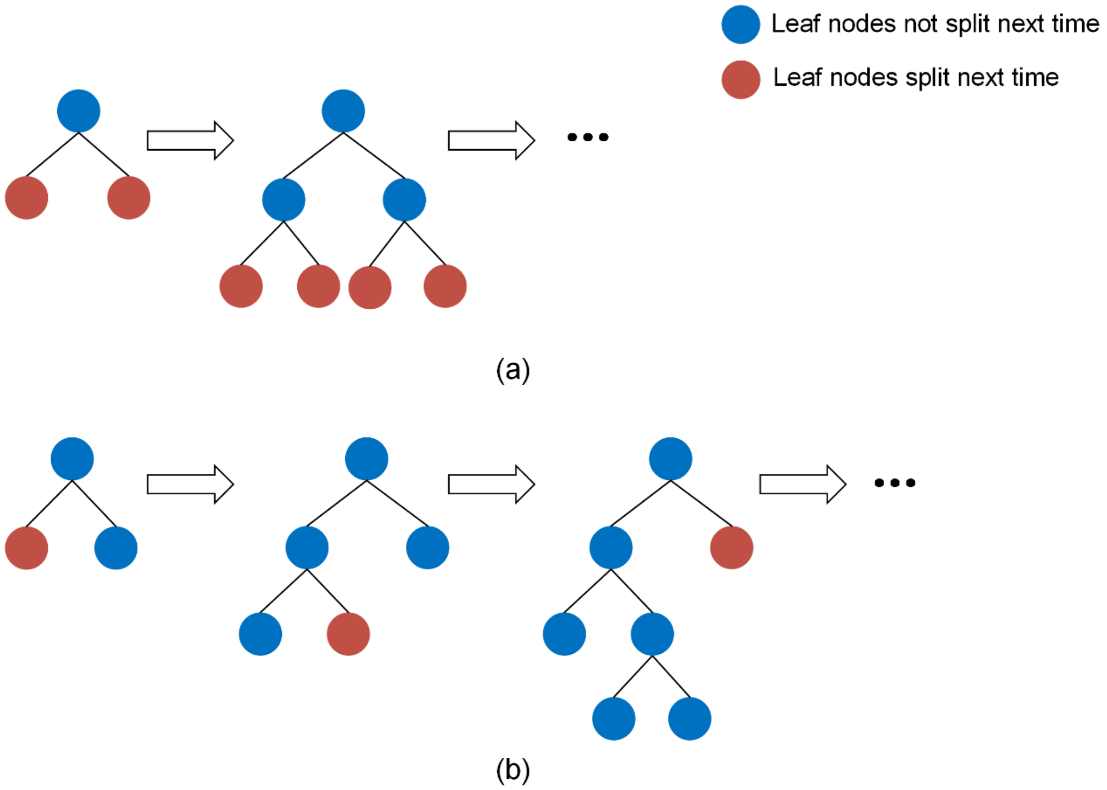


Fig. 6 Schematic diagram of the (a) level-wise and (b) leaf-wise algorithm

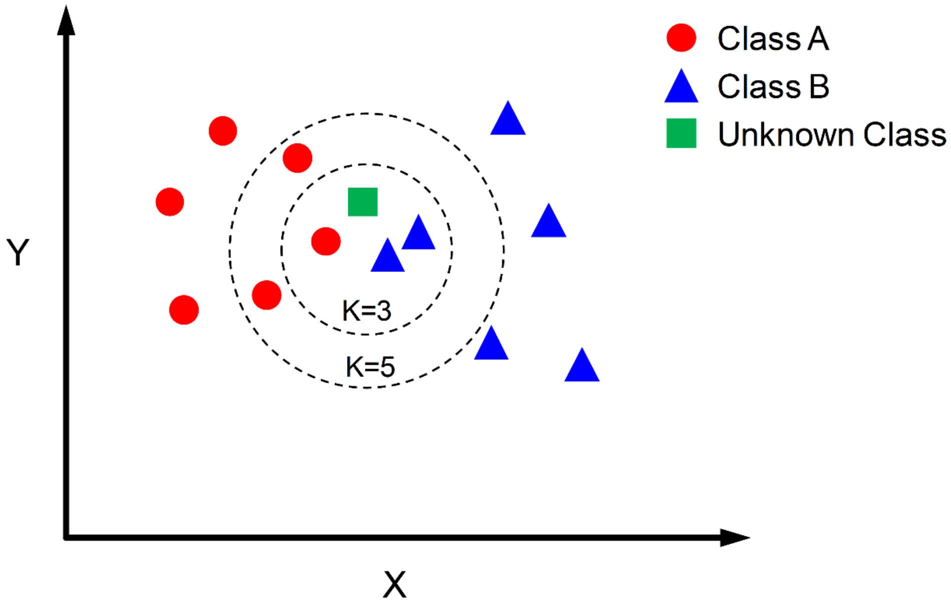


Fig. 7 Schematic diagram of KNN classification

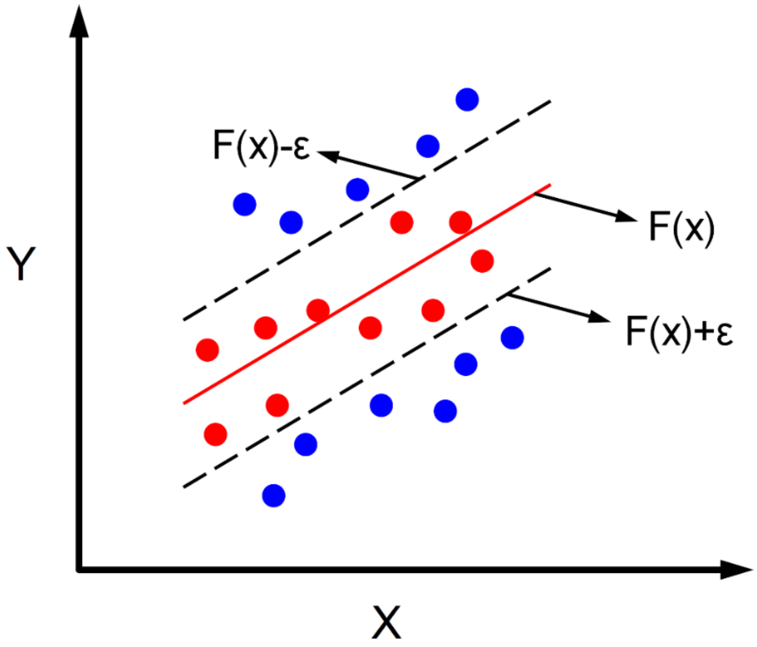


Fig. 8 Schematic diagram of SVR hyperplane data distribution

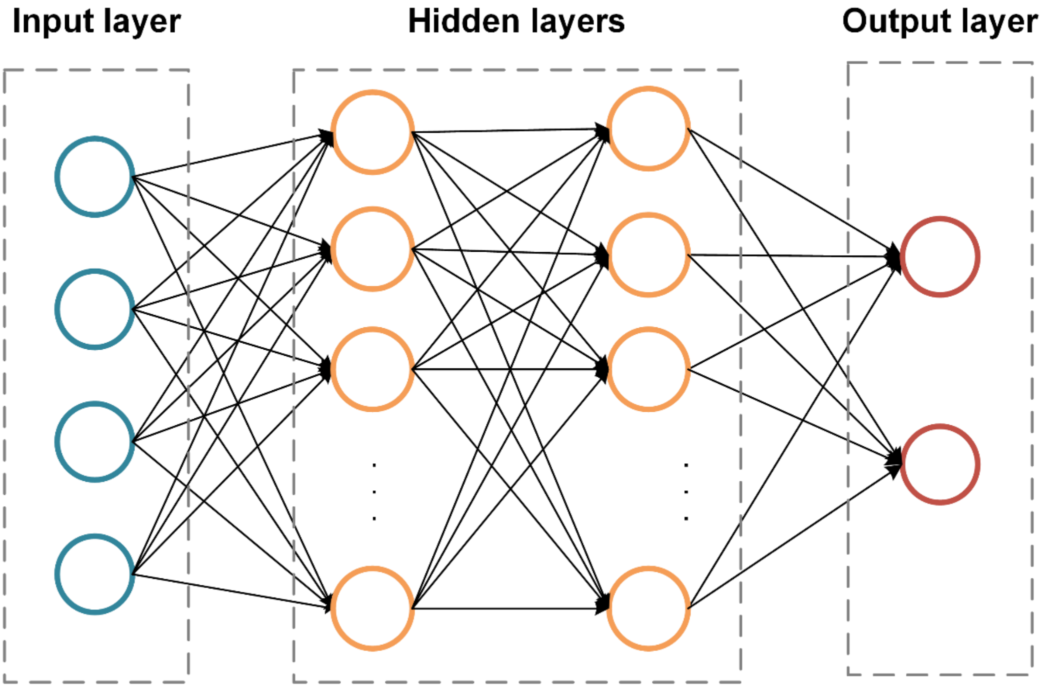


Fig. 9 The structure of BPNN with input layer, hidden layers and output layer

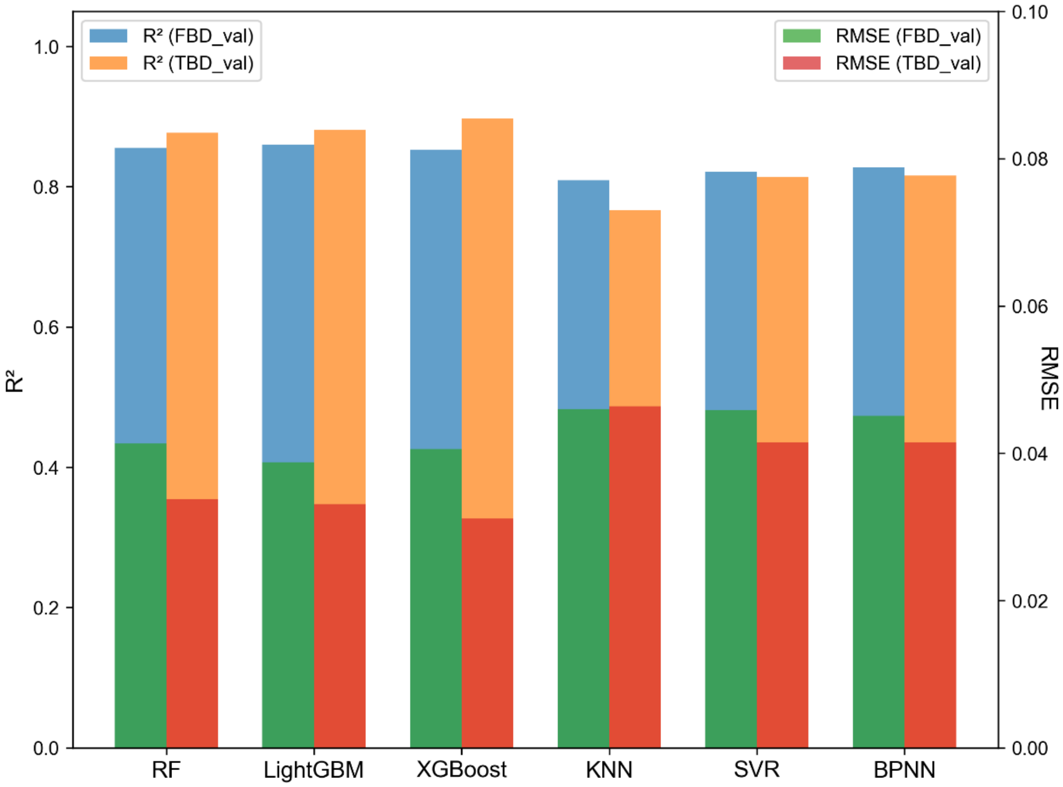


Fig. 10 Performance of the six ML models on validation datasets

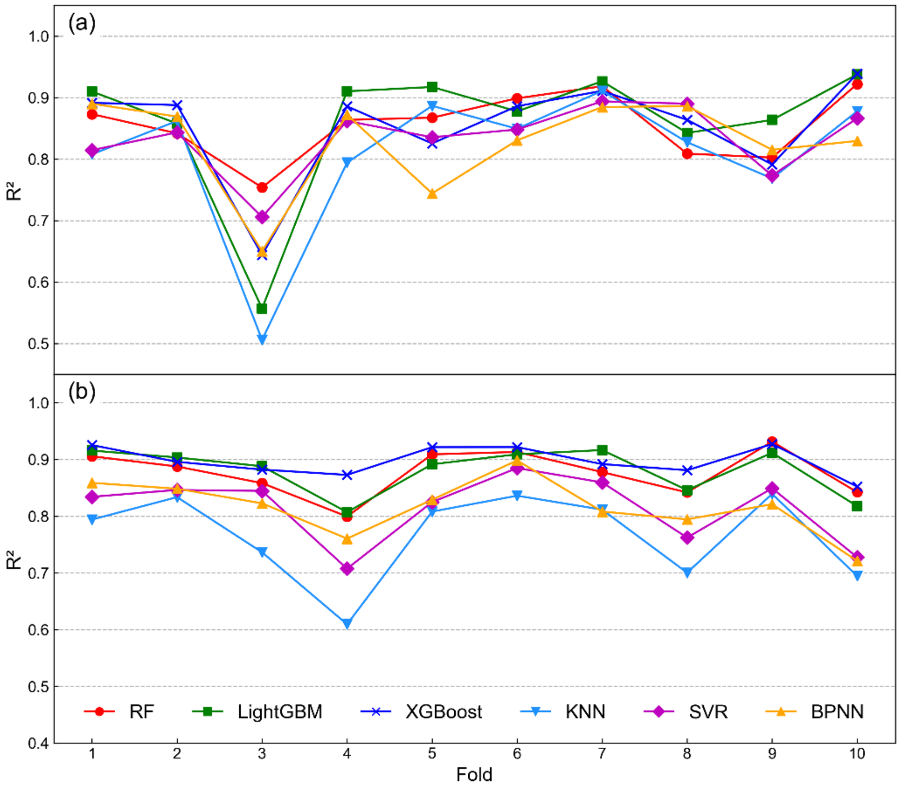


Fig. 11 The R2 results for each fold on (a) FBD_val and (b) TBD_val

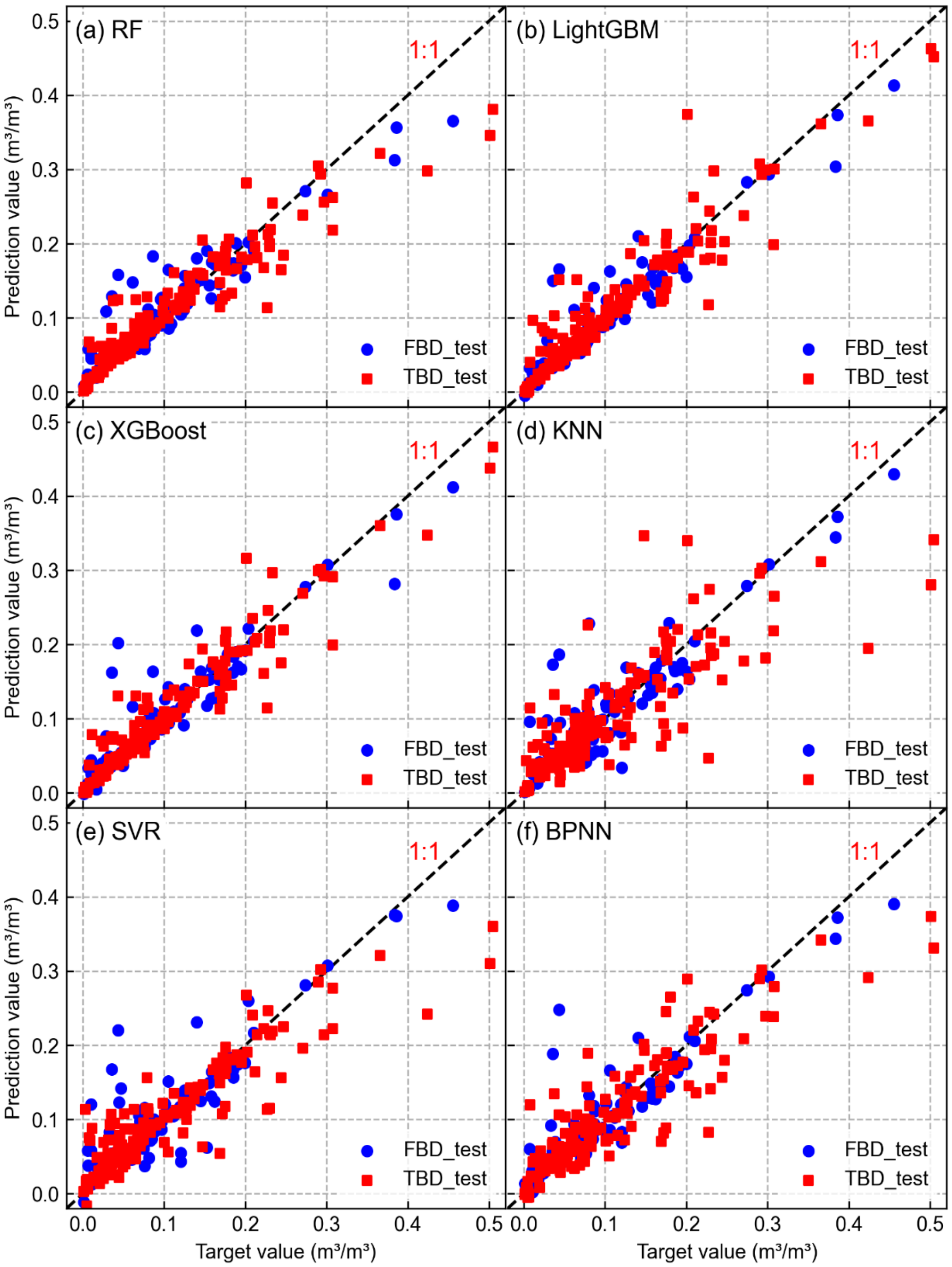
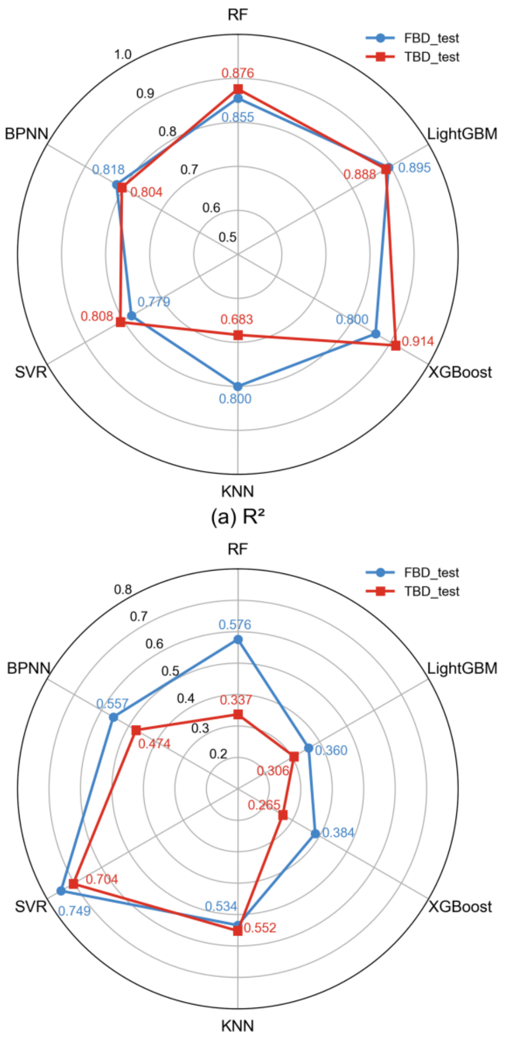


Fig. 12 Prediction results of the six ML models

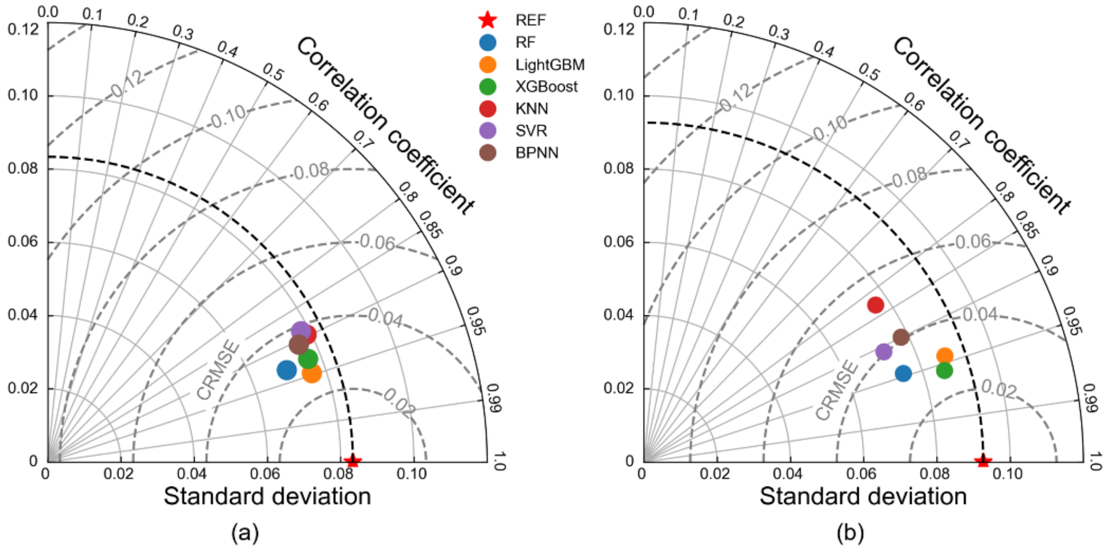


(a) R²

(b) RMSE

(c) MAPE

Fig. 13 Performance comparison of the six ML models



(a)

(b)

Fig. 14 Taylor diagrams of the six ML models on (a) FBD_test and (b) TBD_test