Forest Ecosystems

RESEARCH Open Access

Evaluating soil nutrients of *Dacrydium pectinatum* in China using machine learning techniques

Chunyan Wu[1,2], Yongfu Chen[2*], Xiaojiang Hong[3], Zelin Liu[4] and Changhui Peng[4]

Abstract

Background: The accurate estimation of soil nutrient content is particularly important in view of its impact on plant growth and forest regeneration. In order to investigate soil nutrient content and quality for the natural regeneration of *Dacrydium pectinatum* communities in China, designing advanced and accurate estimation methods is necessary.

Methods: This study uses machine learning techniques created a series of comprehensive and novel models from which to evaluate soil nutrient content. Soil nutrient evaluation methods were built by using six support vector machines and four artificial neural networks.

Results: The generalized regression neural network model was the best artificial neural network evaluation model with the smallest root mean square error (5.1), mean error (− 0.85), and mean square prediction error (29). The accuracy rate of the combined *k*-nearest neighbors (*k*-NN) local support vector machines model (i.e. *k*-nearest neighbors -support vector machine (KNNSVM)) for soil nutrient evaluation was high, comparing to the other five partial support vector machines models investigated. The area under curve value of generalized regression neural network (0.6572) was the highest, and the cross-validation result showed that the generalized regression neural network reached 92.5%.

Conclusions: Both the KNNSVM and generalized regression neural network models can be effectively used to evaluate soil nutrient content and quality grades in conjunction with appropriate model variables. Developing a new feasible evaluation method to assess soil nutrient quality for *Dacrydium pectinatum*, results from this study can be used as a reference for the adaptive management of rare and endangered tree species. This study, however, found some uncertainties in data acquisition and model simulations, which will be investigated in upcoming studies.

Keywords: Support vector machine, KNNSVM, Generalized regression neural network, Nutrient grade, Rare and endangered tree species

Background

Under the conditions of a sharp reduction in global forest area, the speed of species becoming endangered is accelerating and the degradation of forest functions has become serious (Comizzoli and Holt 2014; Sousa-Silva et al. 2014; Comizzoli 2015; Cao et al. 2017; Riccioli et al. 2019). Rare and endangered tree species protection must be strengthened and their growth should be promoted (Comizzoli 2015; Cao et al. 2017). *Dacrydium pectinatum* de Laubenfels (*D. pectinatum*), belonging to *Dacrydium*, genus of the Podocarpaceae family (Farjon and Filer 2013), is a third-class national rare and endangered plant species in the China Red Data Book classification (Fu 1992). It is the only species of this genus that exists in China (Ash 1986; Lian and Yu 2011; Farjon and Filer 2013; Chen et al. 2014). Protection measures must be taken into account for environments that rare and endangered tree species subsist, and abundant soil nutrient is the primary condition for plant subsistence.

* Correspondence: chenyf@caf.ac.cn
[2]Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China
Full list of author information is available at the end of the article

Soil nutrition plays a crucial role in the soil fertility and environmental condition for plant growth and development (Vacca et al. 2017; Camenzind et al. 2018; Chagnon et al. 2018). In addition, many studies have attempted to better quantify and exploit the importance involved in soil nutritional conditions change (Grove et al. 2017; Murphy et al. 2017; Bassaco et al. 2018). Although marked advances have been made in understanding the relationship between soil nutrition and plant growth, researchers remain uncertain about the response of each available soil nutrient as it is related to its content and quality grade, and it has strongly species-specific and differ among congeners. Therefore, the accurate estimation of soil nutrient quality is significant importance to research on the growth of rare and endangered tree species and forest regeneration.

Previous studies have reported several methods used to determine this particular type of estimation accuracy (Zhao et al. 2009). Studies have offered the detail of summaries on the conception and applications, which was used to evaluate soil nutrient quality (Karlen et al. 2001), and they have discussed how to assess soil nutrient quality using field and visible and near-infrared (VNIR) spectroscopy methods. It has provided scientists with an effective method to apply to their research (Gerloff and Krombholz 1966; Idowu et al. 2008). Although this method can obtain accurate data in the field, the major limitation of that is the large amount of sample data and the time it takes to collect samples. In addition, process-based and empirical models have been used to quantify soil nutrient conditions. For example, multiple linear regression (MLR) has been used to predict soil organic stocks in spatial downscaling (Ebrahimi et al. 2017; Roudier et al. 2017). The MLR model provided a unique advantage in simplicity and ease of use (Du 2016; Kawamura et al. 2017). Each of these evaluation methods were suitable and provide a unique advantage. However, they were also subject to robustness conditions (Zhao et al. 2009). In addition, these methods would generate high degrees of error (Zhang et al. 2017). Therefore, determining the best method to evaluate soil nutrition remains a significant challenge.

Currently, machine learning (ML) models have become increasingly popular in agricultural industry and forestry for classification and discrimination (Ghahramani 2015; Shine et al. 2018). ML can improve prediction accuracy (Shine et al. 2018). These include soil microbial dynamic prediction using artificial neural networks (ANN), support vector regression (SVR), and fuzzy inference systems (FIS) (Jha and Ahmad 2018), plant discrimination using support vector machines (SVM) (Akbarzadeh et al. 2018), and soil erosion and nutrient density (Kim and Gilley 2008), soil parameter modeling and classification (Jha and Ahmad 2018), and weed-plant discrimination (Akbarzadeh et al. 2018). In addition, carbon (C), nitrogen (N) and phosphorus (P) content (Li et al. 2017), as well as available N, available P, and nitrate nitrogen (NO_3^--N) (Qi et al. 2018), have already been predicted by ML models (Xu et al. 2015; Moges et al. 2017). These ML models have been demonstrated to have unique advantages in this particular research field. These methods can achieve more accurate estimation results compared to traditional statistical regression methods (Zhao et al. 2009; Zhang et al. 2017). They are powerful tools in coping with small samples, nonlinear relationships without special mathematical equations, and scientific research and practical application hypotheses, especially in the field of high-dimensional pattern recognition (Sun et al. 2016; Zhang et al. 2017), and environments where nutrients are released from agricultural fields (Kim and Gilley 2008; Zhao et al. 2009).

Significant achievements have been made by establishing many advanced and complex algorithmic models. However, these prevalent algorithms are only suitable for a specific plant or communities, but not widely used for other research objectives. In addition, studies on soil nutrient content and quality classification of rare and endangered tree species using ML algorithms are limited (Deng et al. 2017; Moges et al. 2017; Sirsat et al. 2017). A comparative analysis of ML modeling algorithms to determine soil nutrient quality may reduce the difficulty in conducting a quantitative assessment of soil nutrition for the growth of *D. pectinatum* in China, but the application of advanced methods and technology for plants protection and regeneration as well as higher prediction accuracy achievement is an arduous task and a considerable exploratory research endeavor. Therefore, attempting to evaluate soil nutrient content and quality of rare and endangered tree species using ML models is an innovation of this research field. It is also a challenge for ML model application, soil nutrition assessment methods, and tree species research objects.

The main objective of this study was to evaluate the ability of ML algorithms to improve the diagnostic accuracy of soil nutrient quality using soil nutrient data collected from *D. pectinatum*, a vulnerable species in China, formerly dominant in forests in Hainan but excessively logged for more than 20 years. The wood is used in constructing building and ships. The specific aims of the work presented in this study were to: (1) calculate the accuracy of ML models for soil nutrient quality estimations and diagnosis, and (2) determine the optimum ML model and assess model performance.

Methods

Soil nutrition sample preparation

Sample selection

As part of our on-going research effort, three study sites were chosen (Chen et al. 2014) (Fig. 1): the Hainan Bawangling National Nature Reserve (18°57′–19°11′ N,

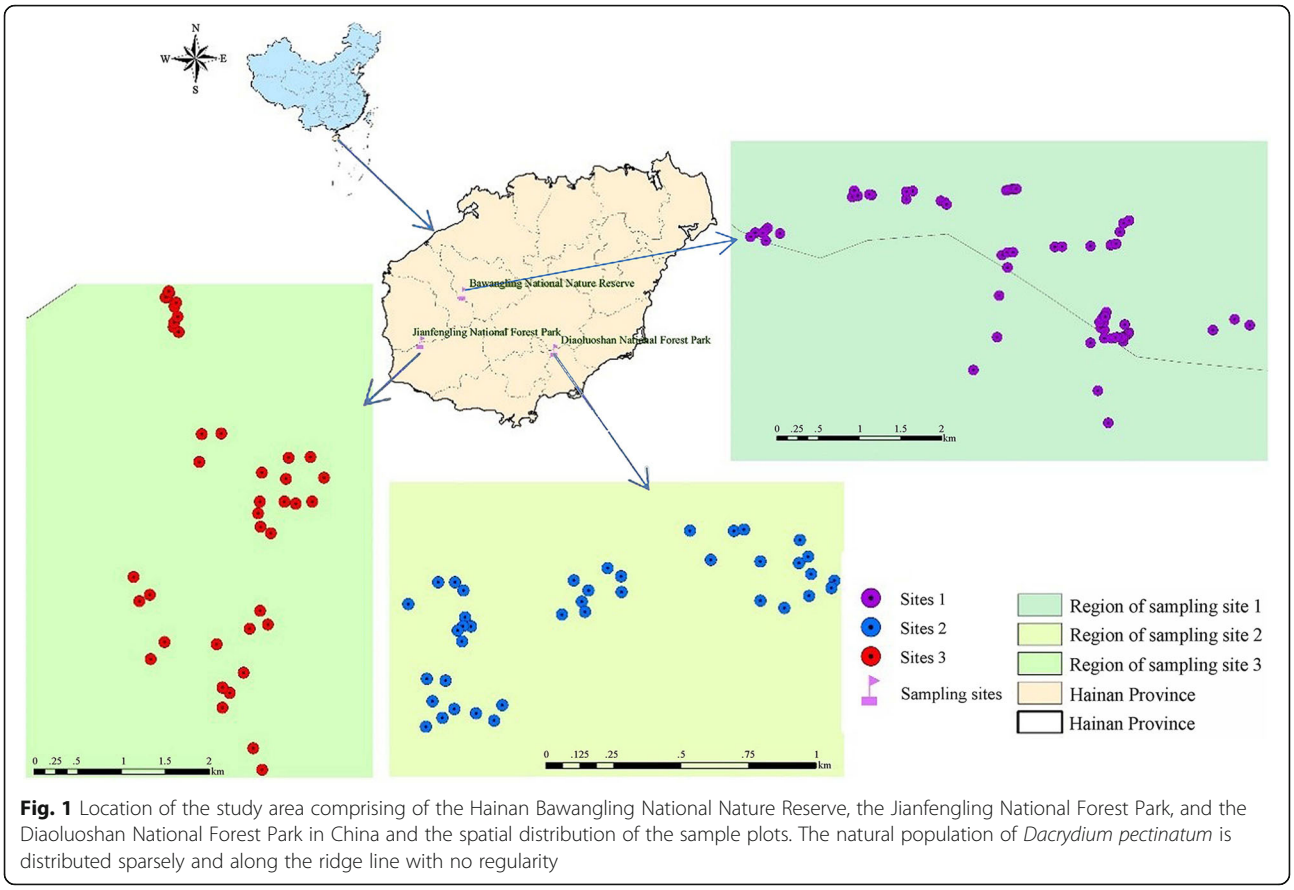

Fig. 1 Location of the study area comprising of the Hainan Bawangling National Nature Reserve, the Jianfengling National Forest Park, and the Diaoluoshan National Forest Park in China and the spatial distribution of the sample plots. The natural population of *Dacrydium pectinatum* is distributed sparsely and along the ridge line with no regularity

109°03′–109°17′ E), the Jianfengling National Nature Reserve (18°24′–18°58′ N, 108°39′–109°24′ E), and the Diaoluoshan National Nature Reserve (18°43′–18°58′ N, 109°43′–110°03′ E), which are the only areas in China where *D. pectinatum* grows. The climate of these three areas is tropical monsoon and tropical sea monsoon. The average annual temperature is approximately 23.6 °C, 19.7 °C, and 24.4 °C, respectively. The average annual precipitation is approximately 1657, 1634, 2400 mm, respectively. The average annual relative humidity of all three sites is above 88%; the forest coverage rate of all three sites is greater than 98%; the altitude of all three sites is approximately 1000–1500 m. These areas are extremely precious as well as being rare, original tropical forests that are under the highest protection priority in China (Lian and Yu 2011). Following a comprehensive survey, a total of 150 experimental plots having seedlings, saplings, and adult trees (Bawangling 72, Jianfengling 38, and Diaoluoshan 40) were selected. A representative typical seedling was used as a center of each plot, with a size of 20 m × 20 m. The plot must be a place where seedlings, saplings and adult trees were concentrated. The number of trees per plot was approximately 8 seedlings, 3 saplings and 1 or 0 adult trees.

Soil sample acquisition

Soil samples were obtained from the three sites in March 2016. These samples were collected from the center point (10 cm beside the center tree stem) and four corners (east, south, west and north corner) point of each plot. Quadrat of soil was 20 cm × 20 cm with a depth of 20 cm in forest plot. Soil samples of 150 g were obtained using a soil auger, for a total of 750 (150 × 5) soil samples, putting them under dry, well-ventilated conditions to allow them to dry naturally and storing them. A wooden hammer was used to break up the dried knots and remove foreign matter (such as plant roots, small stones, glass fragments, etc.) roughly from the soil before being sieved through a 200 mesh. Primary soil samples were then ready, and they were then sent to the laboratory for experimental chemical analysis (Pingree and DeLuca 2018).

Soil nutrient, physical and chemical indicators extraction

Soil nutrient evaluation is essentially a pattern recognition problem, namely, comparing the actual results of the soil nutrient evaluation index system with the corresponding array of soil nutrient evaluation criterion values, which correspond to the array of criterion values closest to outputs array. The soil nutrient quality grade

(output) is the recognition result of the ML model, namely, the soil nutrient evaluation result of the corresponding area. Soil nutrient evaluation cannot be limited to individual nutrient factors. According to previous studies (Wang et al. 2008; Were et al. 2015; Olego et al. 2016), the soil nutrient grading criterion of the second national soil census of China was used as the evaluation criterion (Table 1), using SOM, total N, alkali-hydrolyzable N, available P, and rapidly available K as evaluation indicators. Among the criterion, grade I denotes that soil nutrients are extremely rich and highly concentrated, and much of the content of each nutrient index remains available for most plant growth conditions; grade II denotes that soil nutrients are extremely rich and highly concentrated, and the content of each nutrient index can fully meet the growth needs of plants with a small amount remaining; grade III denotes that the degree of richness and concentration of soil nutrients are within a medium level of availability, and the content of each nutrient index can exactly meet the growth needs of plants (i.e. no surplus); grade IV denotes that soil nutrients are relatively poor and in short supply, and the content of each nutrient index either can meet or not fully satisfy the growth needs of plants; grade V denotes that soil nutrients and supplies are poor, and the content of each nutrient index cannot meet the growth needs of plants; grade VI denotes that soil nutrient availability is extremely poor and in very short supply, and plants are unable to grow under conditions of this nutrient index content.

To obtain inputs of the models, soil samples were chemically analyzed. We measured organic matter using the potassium dichromate volumetric "heating" method (Marcos et al. 2016), the semi-micro Kjeldahl method for the determination of total N (Marcos et al. 2016), and the Kang Hui dish method for alkali-hydrolyzable hydrolysis N content (Marcos et al. 2016). Additionally, we used the 0.5 $mol{\cdot}L^{-1}$ sodium bicarbonate extraction molybdenum-antimony resistance colorimetric method to determine available P (Marcos et al. 2016). We used ammonium acetate in atomic absorption spectrometry to determine the content of rapidly available K (Marcos et al. 2016; Pingree and DeLuca 2018).

Model development and application

This study used soil nutrient grading criterion to measure soil nutrient content and quality through ML modeling. The 10 ML algorithms used in this study was shown in Table 2. The following subsections provide a brief description and implementation details of these 10 methods.

The model consists of an input layer, an output layer, and a hidden layer. The transfer function is a Sigmoid type function that can implement arbitrary nonlinear mapping between input and output, because the ReLU unit would irreversibly die during training, resulting in the loss of data diversification in this study. The collected data were randomly divided into training samples (70%), validation samples (15%) and test samples (15%). Comparing the actual monitoring results of the soil nutrient evaluation index system with the corresponding array of soil nutrient evaluation criterion values, the soil nutrient level corresponding to the array of criterion values closest to the array of monitored values is the recognition result of the artificial neural network model, that is, the results of soil nutrient evaluation in the corresponding area (Fig. 2). In this study, N, organic matter content, alkali-hydrolyzable N, available P, and rapidly available K were used as inputs, and soil nutrient quality grades were used as outputs.

Artificial neural network

ANN is an artificial intelligence technology that has been developed in recent years to simulate biological processes of the human brain (Guo et al. 2017). ANN model analyzes the internal relationships and regular patterns of two variables by providing a set of mutually corresponding input and output data, and then forms a complex nonlinear system function through these regular patterns (Zhao et al. 2009). Our study attempted to get the outputs using a series of ANN models, applying the pattern recognition function of the ANN model. The BPNN algorithm adjusts weight and deviation values along a negative gradient to attempt to minimize the mean squared error (MSE) of the input and output

Table 1 Soil nutrient content evaluation criteria

Grade	I	II	III	IV	V	VI
State	Extremely high	High	Medium	Low	Lower	Extremely low
SOM ($g{\cdot}kg^{-1}$)	> 40	30–40	20–30	10–20	6–10	< 6
N ($g{\cdot}kg^{-1}$)	> 2.0	1.5–2.0	1.0–1.5	0.7–1.0	0.5–0.7	< 0.5
AHN ($mg{\cdot}kg^{-1}$)	> 150	120–150	90–120	60–90	30–60	< 30
AP ($mg{\cdot}kg^{-1}$)	> 40	20–40	10–20	5–10	3–5	< 3
RAP ($mg{\cdot}kg^{-1}$)	> 200	150–200	100–150	50–100	30–50	< 30
P ($g{\cdot}kg^{-1}$)	> 1	0.8–1.0	0.6–0.8	0.4–0.6	0.2–0.4	0.0–0.2
K ($g{\cdot}kg^{-1}$)	> 25	20–25	15–20	10–15	5–10	0–5

Note: SOM refer to soil organic matter; N refer to total nitrogen; AHN refer to alkali-hydrolysable nitrogen; AP refer to available phosphorus; RAP refer to rapidly available potassium; P refer to phosphorus; K refer to potassium

Table 2 The detail of abbreviations and full name of 10 methods

Abbreviations	Full name
BPNN	Back-propagation neural network
FPNN	Field probing neural network
MLPNN	Multilayer-propagation feed forward neural network
GRNN	Generalized regression neural network
LMSVM	Local mixture-based support vector machine
SVM-KNN	*k*-NN and SVM integrated algorithm
KNNSVM	*k*-nearest neighbors local support vector machine
LSVM	Localized support vector machine
PSVM	Proximal support vector machine
FSVM	Fast local kernel support vector machine

(Zhao et al. 2009). The benefit of the FPNN algorithm is that its calculation is elemental, it has a simple network structure, and its learning complexity is minimal. MLPNN is the most time-saving method. GRNN has a strong nonlinear mapping capacity and requires a small sample size (Myers et al. 2017). ANNs can predict and evaluate a network more quickly and provide greater computational advantages as the advantages of a fast convergence rate, high prediction accuracy, fewer adjustment parameters, and not being easy to fall into local minima.

Partial support vector machine

SVMs are commonly used in classification and recognition due to the small number of training samples required and the high accuracy of results (Gunn 1998). However, the classification accuracy of SVM is also affected by issues such as the selection and optimization of kernel functions, the establishment of multiple SVM models, and type selection and soil nutrient characteristics (Shu 2015). In order to further improve the classification effect of SVM, soil nutrients were classified using multiple partial SVM methods (Brailovsky et al. 1999), which combine *k*-nearest neighbors (*k*-NN) with SVM. Partial SVM methods are divided into six categories.

Category I: The local mixture-based SVM (LMSVM) can realize the locality of SVM by adding two multipliers to the kernel function (Eq. 1).

$$\sum_{r=1}^{n_w} K(x_i, x_j) h(|x_i, w_r|) h(|x_j, w_r|) = n_w K(x_i, x_j) \quad (1)$$

where $r = 1, 2, 3,..., n$; n_w is the sample number that the sample is concentrated to meet $|x_i, w_r| \le \theta_r$; and $K(x_i, x_j)$ is the prokaryotic function. In this research, n is 300.

Category II: The *k*-NN and SVM integrated algorithm (SVM-KNN). Firstly, *k*-nearest neighbors are searched for the unlabeled sample *x*', and then the distance between the unlabeled sample and the *k*-nearest neighbors is calculated to form a distance matrix. The distance matrix is directly transformed into a kernel matrix. Finally, the class label of the unlabeled sample is determined by using DAGSVM algorithm.

Category III: The *k*-NN local support vector machine (KNNSVM) algorithm calculates the distance in kernel space and avoids the instability caused by the nonlinear issue of the different classifications (Shu 2015). The algorithm finds the *k*-NN for each unlabeled sample in the training set. By using *k*-NN to establish a SVM classifier, the class label of the unlabeled samples can be obtained.

Category IV: The localized support vector machine (LSVM), which establishes the similarity factor between unlabeled samples and training samples, and adds penalties for SVM constraints (Cheng et al. 2010). Optimization issues related to LSVM are as follows:

$$\begin{cases} \min \dfrac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N} \delta(x' + x_i)\xi_i \\ s.t. y_i(\omega^T x_i + b) \ge 1 - \xi_i \\ \xi_i \ge 0, i = 1, 2, ..., N \end{cases} \quad (2)$$

where $\delta(x' + x_i)$ is the similarity factor; C is the constant.

When the value of $\delta(x' + x_i)$ is [0, 1], then

$$\delta(x' + x_i) = \exp\left(\frac{-\|x' - x_i\|^2}{2\delta^2}\right) \quad (3)$$

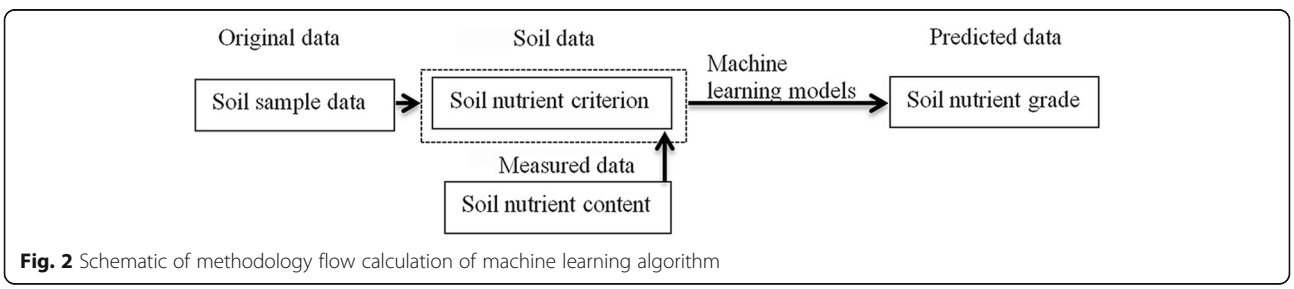

Fig. 2 Schematic of methodology flow calculation of machine learning algorithm

Category V: Establishing SVM for each cluster center by clustering the training samples, and then using these SVMs to classify unlabeled samples, namely, PSVM (Hao 2016), the clustering function is:

$$\min_{C,Z} \sum_{j=1}^{k} \sum_{i=1}^{n} Z_{i,j} \left\| x_i - C_j \right\|_2^2 + R \sum_{j=1}^{k} \left| \sum_{i=1}^{n} Z_{i,j} y_i \right| \quad (4)$$

where y_i is the class label of the i^{th} training sample, and x_i is the i^{th} row in the similarity matrix of the unlabeled sample and the training sample. C_j is the j^{th} cluster center; R is a scaling parameter; and $Z_{i,j}$ is an element in the inverse of the cluster.

Category VI: The fast local kernel support vector machine (FSVM), which improves the method of solving the cluster center point (Segata and Blanzieri 2010).

Model training and validation data

The basic structure of the ANN model in estimating soil nutrient quality grades (Giovanis et al. 2017) was shown in Fig. 3.

The input layer was the soil nutrient content by field measurements, and the output layer was the soil nutrient grades. In addition, given that the basic principle of SVM (Li et al. 2014) was a quadratic algorithm to determine the best hyper plane, it was shown in Fig. 4, and all samples were separated from the maximum interval boundary (Cristianini and Shawe-Taylor 2000). The 10 above mentioned algorithms were used as the soil nutrient evaluation models in this study.

The characteristics of soil nutrient quality were determined by soil nutrient content and the cross-fertilization characteristics between the nutrients themselves. The input variables selection of ANN model was determined by the soil nutrient grade criteria.

For the BPNN, we have determined the input variables that were total N, organic matter content, alkali-hydrolysable N, available P, and rapidly available K were selected. The hierarchical structure establishment for this model was that the number of nodes in the input layer was 5. Taking soil nutrient quality grade as output unit, that is, the number of nodes contained in the output neuron unit was 1, the initial number of nodes in the hidden layer was set to 11. The weight learning function uses the trainlm algorithm, using the non-linear continuous derivable excitation function, and the node's transfer function is purelin. Specific model parameters were shown in Table 3.

And then, BPNN was established with three-layer. The BP network was trained by inputting all the data sets as samples. In the training process, the above original data are normalized by using Premnmx function in MATLAB, so that the data set is between [− 1, 1].

For the FPNN, we selected the total N, organic matter content, alkali-hydrolysable N, available P, and rapidly available K as input variables. The training and calculation process of the model was that when P samples are given, a total of P–1 elements are taken from the first layer. Each element has n inputs (assuming that the input of the sample is n dimension and the output is m dimension). The function of this layer element is to transform the input of P samples into p vertices of orthogonal p–1 dimensional simplex in P–1 dimension space. From the second layer to the third layer, take M components. Through the second layer element, the p

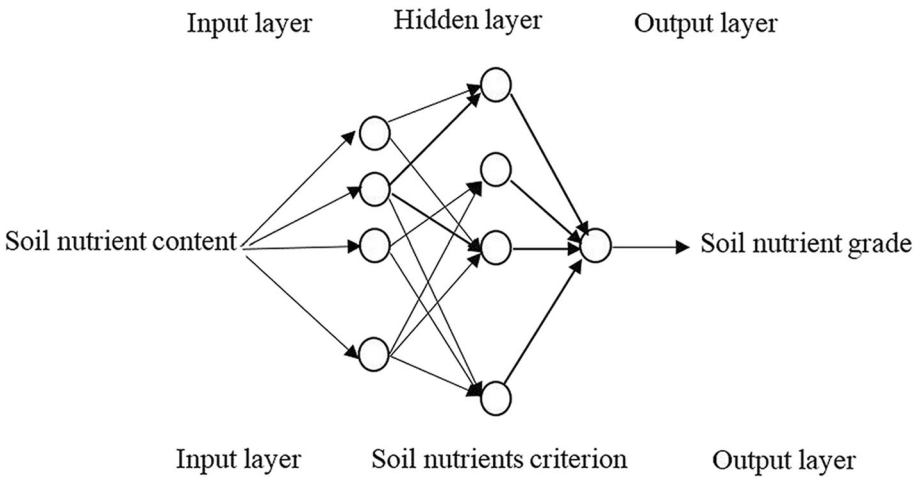

Fig. 3 A schematic structure of a neural network. Field measurements taken from soil nutrients (total nitrogen, organic matter content, alkali-hydrolysable nitrogen, available phosphorus, and rapidly available potassium) were used as input values provided through the input layer. Five evaluation indicators were used as dependent variables. Soil nutrient quality evaluation results are represented by an output node. Four layers of a hidden layer were constructed

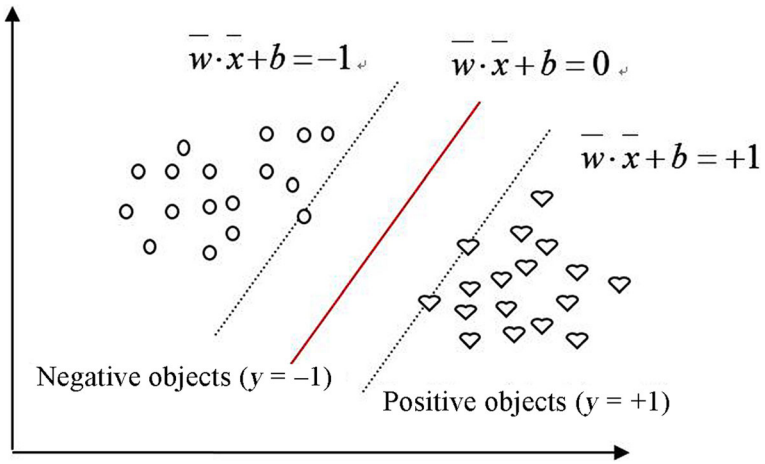

Fig. 4 Sample points on the hyper plane (also referred to as support vectors). Obviously, if $f(x) = 0$, then x is the point on the hyper plane. It may be required that for all points that satisfy $f(x) < 0$, the corresponding y is equal to $y = -1$, and $f(x) > 0$ corresponds to data points where $y = 1$. wx denotes the inner product of w and x in the classification function $f(x) = wx + b$. In the common sense, w is the normal vector (weight vector) and b is the intercept (bias)

vertices of the orthogonal p–1 dimensional simplex are transformed into $p \times m$ dimension sample output vectors, and then the neural network corresponding to the associative memory of the sample set is obtained.

For the MLPNN, let K be given an input set $K = \{x^1, x^2, \ldots, x^k\}$ (k is the point set of n-dimensional Euclidean space), let K be divided into s subsets $K^1 = \{x^1, x^2, \ldots, x^{m(1)}\}$, …, $K^s = \{x^{m(s-1)+1}, \ldots, x^k\}$. This paper presents a three-layer network N, which satisfies: the output of points belonging to K^i is "y^i" after passing through this network, where $y^i = (0, \ldots, 1, 0, \ldots, 0)$ (i.e. the vector whose first component is 1 and the rest component is 0). $i = 1, 2, \ldots, s$.

Table 3 The architecture of BPNN and model parameter selection for weight learning function and node transfer function

Model parameter	Parameter quantity
Training function	Tranilm
Learning function	Learndm
Performance function	MSE
Hidden layer transfer function	Tansig
Output layer transfer function	Purelin
Number of neural elements nodes in input layer	5
Number of nodes in output layer neural units	1
Learning rate	0.4
Momentum coefficient	0.8
Iteration times	≤ 50,000
Network convergence error	≤ 0.05
Inertia factor	0.5
Training target error	0.001
Initial weight	[0 + 0.5]
Learning coefficient	0.05

The main idea of this algorithm is to find a field C^1, which covers only the points in K^1 but not those in K^2, then delete the points covered by C^1 and find another field C^2 for the remaining points. It covers only the points in K^2 but not those in K^1, and then deleted the points covered by C^2, so that the points covered by C^2 are intersected and covered until all the points in K^1 (or K^2) are deleted. The specific steps are 1) Mapping the points of K^1 and K^2 onto the spherical surface S^n (S^n is $n + 1$ dimensional space, the center is at the origin, the radius is equal to the n dimensional sphere of R, and the radius $R > \mathrm{Max} \mid X^i_j \mid$) is still recorded as K^1, K^2. 2) Find the initial point and cover it from that point. Calculate the center of all samples and find the nearest sample point a. 3) Determine the radius of the covering area C^1 centered on a. Find the nearest dissimilar point b, whose distance is denoted d as d_1, and then find the farthest similar point c whose distance is less than d, whose distance is denoted as d_2, then the radius of coverage area $r = d_1 + d_2/2$. 4) Focus on domain C^1. 5) Repeat steps 3) and 4) until the number of samples covered is not more than that before the center of gravity. 6) Repeat steps 3) to 5) until the number of samples covered is not more than the number of samples covered in the previous one. A local maximum field C^1 covering K^1 points is obtained. The subset of K^1 covered is marked as K^{1i}. 7) Find a different point to start covering. Its category is K^2. Let $T \le K^1/K^{1i}$, $K^1 \le K^2$, $K^2 \le T$. 8) Repeat steps 3) to 6) until only the last category is left. 9) Processing the last class of points to get the last coverage.

For the GRNN, we selected the total N, organic matter content, alkali-hydrolysable N, available P, and rapidly

available K as input variables. In the current study, an iterative process using quad cross-validation is utilized to determine the optimal smoothing factor according to our experience and other applied research results, and this factor ranges from 0.01 to 1 (Dou and Yang 2017). The probability density function used in GRNN is the normal distribution. The function is

$$SNC = \frac{\sum SNC_i\, k(x,\ x_k)}{\sum k(x,\ x_k)} \tag{5}$$

where SNC is soil nutrient content, x is the input that is the total N, organic matter content, alkali-hydrolysable N, available P, and rapidly available K, SNC_i is the activation weight for the pattern layer neuron at k, $k(x,\ x_k) = e^{-d_k/2\sigma^2}$, $d_k = (x - x_k)(x - x_k)^T$, where d_k is the squared Euclidean distance between the training samples x_k and the input x. The steps for the calculations in this study include 1) calculating distances d_1, d_2, …, d_k (1.04, 0.86, 1.74, 0.74, 1.07); 2) calculating weights using the activation function $e^{-d_k/2\sigma^2}$; 3) summing w's, $W = w_1 + \ldots + w_k$ and the numerator was $f(x)$ $w = w_1 SNC_1 + \ldots + w_k SNC_k$; and 4) calculating the predicted output $SNC(w/W)$.

The BPNN, FPNN, MLPNN, and GRNN algorithms (Cheng 2005) were used to establish the soil nutrient evaluation model for comparison. In this study, four ANN algorithms were used to estimate soil nutrient content and quality. In order to resolve the over-fit caused by hyper-parameters during adjustment and the calculation of model accuracy, the hyper-parameters must be minimized by adopting a hierarchical nested cross validation method. All models were trained more than 1000 times each time. The training-set, validation-set, and testing-set for each time were different for each model.

Soil field measurement data were used to calibrate and validate the model. Soil nutrition reference data were randomly divided in to a calibration set and a validation set. After experimenting with several time calculations, it was determined that we obtained the best result when 70%, 15%, and 15% of the dataset were used as a training, validation, and test set, respectively. This was in agreement with a similar previous study (Li et al. 2014). The calibration set is used to train ANN models and validation sets by validating ANN model performance. Cross-validation is used to test the models.

For the KNNSVM, SVM-KNN, FSVM, PSVM, LSVM, and LMSVM, the algorithm steps are 1) According to a certain principle, the samples in the training set are divided into k classes and K centers are found. 2) Clustering each training sample using K-means, generating n sample centers instead of the original training samples, and constructing a support vector machine for each center. 3) Find a center closest to x for each test sample. 4) Use the support vector machine corresponding to the center to name x. 5) Outputs the results. For the SVM models, 30%, 25%, 35%, 20%, 30%, and 20% of the data were selected randomly as a testing set, and 70%, 75%, 65%, 80%, 70%, and 80% of the data were chosen as the training set, respectively. The selected training and testing samples were feature extracted and normalized, and they formed an eigenvector matrix of the entered training data. Each model was calculated 1000 times under different conditions to provide for more accurate predictions.

Soil nutrient quality was determined though a soil nutrient assessment and described in grades. The outputs, namely, extremely high, high, medium high, low, poor, extremely poor, were respectively recorded as a range instead of a fixed value. The prediction performance of the model was evaluated according to the calculated mean error (ME), the mean square prediction error (MSPE), and the root-mean-squared error (RMSE) (the square root of the MSPE) (Bibby and Toutenburg 1977; Shine et al. 2018). In addition, MATLAB software (version 8.2, The MathWorks, Inc., Natick, MA, USA) and its toolboxes were utilized to analyze the data in this study.

Results

Model performance

The soil nutrient content of the experimental sites was estimated by four ANN algorithms. The RMSE, ME and MSPE indices were used in order to evaluate the efficiency of the four models. RMSE, ME, and MSPE with smaller values indicated higher model efficiency. The result was shown in Table 4.

The GRNN model was best at evaluating the ANN models with a RMSE of 5.1 (Table 4). The MSPE of the GRNN model was 29, and the ME was − 0.85. According to the averages of the RMSE, MSPE, and ME in the table, the GRNN model was determined to be the best ANN model to estimate soil nutrient content.

The improved six partial SVM models were used to classify the training and testing samples for the selected outputs. The six partial SVM models were used to classify soil nutrient elements in the training set samples with an accuracy rate greater than 90%. Soil nutrient grade was tested on the samples (Table 5). Among the six SVM models, the KNNSVM model had the highest accuracy rate (93.6%), followed by the SVM-KNN model (91.9%), the FSVM model (89.9%), the PSVM model (88.9%), the LSVM model (86.8%), and the LMSVM model (85.4%). The results showed that the improved partial SVM models were suitable in improving the accuracy of soil nutrient evaluation.

As shown in Fig. 5, the average prediction accuracy of all four ANN models was greater than 88%. The GRNN model yielded the highest accuracy value (92.5%) among the four models. The accuracy of the MLPNN model

Table 4 Artificial neural network calculation results

Algorithm	Training set	Testing set	RMSE	MSPE	ME
BPNN	210	45	5.6	31	−0.59
FPNN	210	45	6.8	42	−0.53
MLPNN	210	45	7.2	55	−0.51
GRNN	210	45	5.1	29	−0.85

Note: BPNN is the back-propagation neural network; FPNN is the field probing neural network; MLPNN is the multilayer-propagation feed forward neural network; GRNN is the generalized regression neural network; MSPE is the mean square prediction error; RMSE is the root-mean-squared error; ME is the mean error

(88.5%) was relatively low. The results also showed that the GRNN model was slightly more stable than the others, while the stability of the BPNN model was ranked second and the FPNN model was ranked third.

The receiver operating characteristic (ROC) curve was used to evaluate the performance of the modeling classifier (Guo et al. 2017). Figure 6 shows the ROC curves for the four models, which illustrates the performance of a classification model under all classification thresholds and depicts two parameters: the false positive rate (FPR) represented by the *x*-axis, and the true positive rate (TPR) represented by the *y*-axis. The area under the curve (AUC) represents classification performance, which is the ability of the target model to correctly classify the different outputs. The AUC value (0.6572) of the GRNN model was the highest of the four models, followed by the BPNN model with an AUC value of 0.6486, and the FPNN model with an AUC value of 0.6475. The AUC value of the MLPNN model was the lowest at 0.6459.

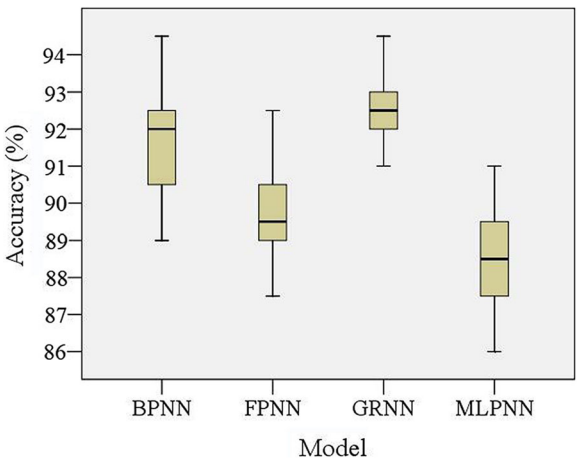

Fig. 5 Evaluation results of the different models. BPNN is the back-propagation neural network; FPNN is the field probing neural network; GRNN is the generalized regression neural network; MLPNN is the multilayer-propagation feed forward neural network

After a comprehensive evaluation, this study determined that there were six soil nutrient grades. Moreover, 210 training samples were used to train the neural network. The training step number was 1000, and the four ANN models were cross-validated. The results are shown in Table 6.

Four models were verified by a mixed matrix (Table 6). The sum of diagonal prediction values of each sub-table in the table of each model was the prediction accuracy of the cross-validation. All models had good prediction accuracy, namely, greater than 87%. The prediction accuracy of the GRNN model reached 92.5%, but showed no

Table 5 Accuracy of the sample training set and testing set of the six models

Soil nutrient grade		KNNSVM (%)	LSVM (%)	LMSVM (%)	SVM-*k*-NN (%)	FSVM (%)	PSVM (%)
Training set ratio	I	96.9	96.5	80.4	92.8	81.7	88.2
	II	97.5	95.8	90.7	93.4	80.8	89.7
	III	96.1	92.9	93.2	92.8	98.3	94.8
	IV	98.5	85.8	98.9	98.0	98.7	92.4
	V	96.0	88.3	80.1	93.3	97.6	95.2
	VI	96.6	91.1	98.8	91.7	96.3	93.8
	Average	96.6	91.7	90.4	93.7	92.9	92.4
Testing set ratio	I	97.0	83.6	87.5	81.1	91.1	92.4
	II	90.0	87.4	90.7	86.5	92.7	90.7
	III	95.5	91.3	88.6	98.2	91.7	93.3
	IV	86.8	82.9	80.4	97.7	87.5	89.1
	V	94.4	87.3	81.3	90.2	88.6	90.4
	VI	98.1	88.5	84.1	97.7	89.7	80.6
	Average	93.6	86.8	85.4	91.9	89.9	88.9

Note: KNNSVM is the *k*-nearest neighbors local support vector machine; LSVM is the localized support vector machine; LMSVM is the local mixture-based support vector machine; SVM-KNN is the *k*-NN and SVM integrated algorithm; *FSVM* is the fast local kernel support vector machine; PSVM is the proximal support vector machine. I, II, III, IV, V, and VI indicate that the soil nutrient quality grade is "extremely high", "medium high", "low", "poor", and "extremely poor", respectively

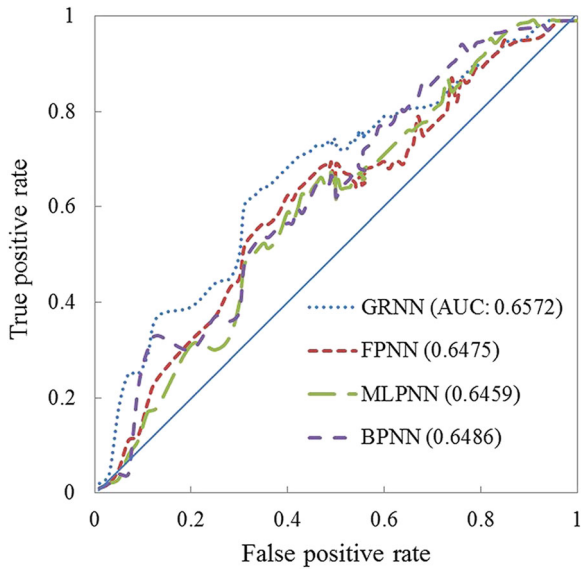

Fig. 6 Box-whisker plot of accuracy. GRNN is the generalized regression neural networks; FPNN is the field probing neural network; MLPNN is the multilayer-propagation feed forward neural network; BPNN is the back-propagation neural network. AUC is the area under the curve. The AUC represents classification performance, which is the ability of the target model to correctly classify the different outputs

significant difference. The accuracy of the four ANN models was 92.5%, 92.0%, 89.5%, and 88.5% for the BPNN, FPNN, GRNN, and MLPNN models, respectively. The four models had the highest percentages of false positives, namely, 3.8%, 3.9%, 4.2%, and 3.5% for the BPNN, FPNN, GRNN, and MLPNN models and false negatives, namely, 3.9%, 3.6%, 3.7%, and 4.1% for the BPNN, FPNN, GRNN, and MLPNN models, respectively. Overall, the GRNN model had the highest soil nutrient assessment accuracy.

Evaluation accuracy rate of the different models

The accuracy of the output of KNNSVM model was higher than other five partial SVM models (Fig. 7), while the accuracy of the LMSVM model was the lowest among the six SVM models. In addition, the evaluation accuracy of the outputs of the various models also differed. The KNNSVM model had the highest assessment accuracy in soil grade VI (95.1%), V (94.3%), II (93.2%), I (92.0%), III (90.0%), and IV (88.2%). The PSVM model had the highest assessment accuracy in grade VI (91.6%), V (89.4%), IV (88.3%), II (86.0%), III (85.6%), and I (85.1%). The LSVM model had the highest assessment accuracy in grade V (89.5%), IV (88.7%), VI (87.7%), II (86.4%), III (85.8%), and I (84.1%). The LMSVM model had the highest assessment accuracy in grade VI (87.1%), V (86.2%), IV (85.5%), III (84.4%), II (83.7%), and I (83.1%). The SVM-KNN model had the highest assessment accuracy in grade V (93.3%), VI (92.8%), IV (92.6%), II (92.4%), III (88.5%), and I (88.3%). The FSVM model had the highest assessment accuracy in grade V (92.7%), VI (92.1%), IV (88.2%), III (87.5%), II (85.8%), and I (85.6%).

Table 6 Confusion matrix from cross-validation of the four artificial neural network model results

	Grade	Measured grade (%)					
		I	II	III	IV	V	VI
Predicted grade	I	15.4	0.5	1.0	1.7	1.0	0.8
	II	1.2	15.5	2.0	1.2	0.5	0.5
	III	0.8	0.7	15.1	2.0	0.4	1.1
	IV	0.9	1.1	1.7	15.7	0.8	0.6
	V	1.2	1.3	0.9	0.3	15.2	0.5
	VI	0.3	0.4	0.6	0.5	0.8	15.6
Accuracy		92.5					
Predicted grade	I	15.7	1.3	1.5	0.6	0.8	1.6
	II	0.4	15.5	1.2	0.7	0.9	0.3
	III	0.3	2.1	15.2	0.9	1.3	0.4
	IV	0.3	1.5	1.8	15.3	0.4	0.8
	V	0.7	0.8	1.2	0.8	15.4	0.5
	VI	0.5	0.7	0.7	0.6	0.4	14.9
Accuracy		92.0					
Predicted grade	I	15.7	0.8	1.1	0.7	0.9	0.5
	II	1.7	14.8	0.7	0.5	0.6	1.4
	III	0.6	0.7	14.4	0.6	1.8	1.1
	IV	0.3	0.6	0.3	15.5	1.2	1.6
	V	0.8	1.7	0.2	0.9	14.7	1.4
	VI	0.9	1.5	0.5	0.4	0.6	14.4
Accuracy		89.5					
Predicted grade	I	14.7	0.6	1.2	0.9	1.1	0.7
	II	1.4	14.4	0.5	0.7	0.8	1.6
	III	0.7	0.9	15.0	0.7	1.3	1.0
	IV	0.8	0.8	0.6	14.9	1.6	1.4
	V	0.5	1.3	0.3	0.6	15.2	1.3
	VI	0.7	1.7	0.4	0.9	0.7	14.3
Accuracy		88.5					

Note: I, II, III, IV, V, and VI indicate that the soil nutrient quality grade is "extremely high", "medium high", "low", "poor", and "extremely poor", respectively. Four models are back propagation neural network (BPNN), field probing neural networks (FPNN), muhilayer perceptron neural networks (MLPNN), and general regression neural network (GRNN)

Discussion

This study determined the best model to predict and evaluate soil nutrition by investigating the adaptability and validity of a variety of ML techniques with data from areas where a rare and endangered tree species, *D. pectinatum*. The four ANN and six SVM models, namely, the

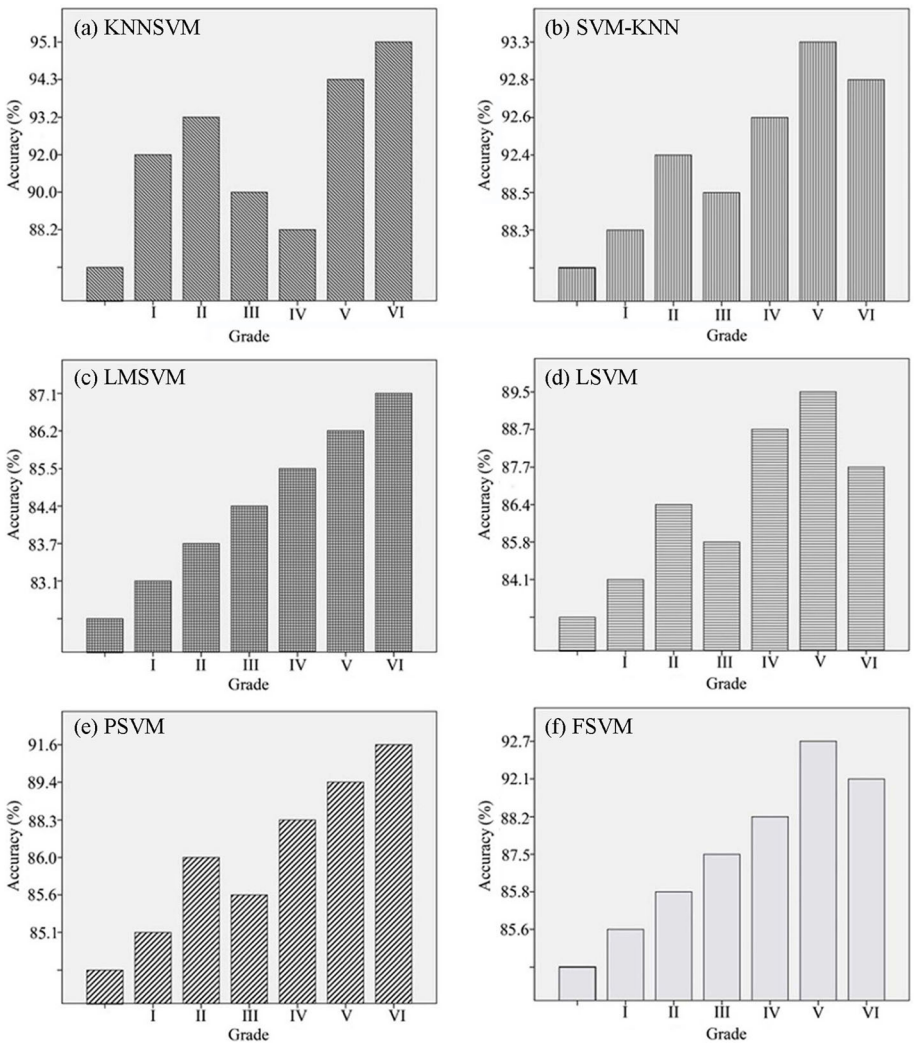

Fig. 7 Assessment accuracy of the six models. **a** KNNSVM is the *k*-nearest neighbor local support vector machine; **b** SVM-KNN is the *k*-NN and SVM integrated algorithm; **c** LMSVM is the local mixture-based support vector machine; **d** LSVM is the localized support vector machine; **e** PSVM is the proximal support vector machine; **f** FSVM is the fast local kernel support vector machine. I, II, III, IV, V, and VI indicate that the soil nutrient quality grade is "extremely high", "medium high", "low", "poor", and "extremely poor", respectively

BPNN, FPNN, MLPNN, GRNN, KNNSVM, LSVM, LMSVM, SVM-KNN, FSVM, and PSVM models, were used to get the outputs. This study used RMSE, MSPE, and ME to determine prediction performance efficiency of the models. In addition, in order to obtain accurate outputs, soil nutrient grading criteria based on previous studies were used as the estimation criteria. This study also used cross-validation to obtain a high prediction accuracy of the ANN models. Following this, the evaluation accuracy rate and ROC of the SVM models were used to establish the main nutrient content to determine the best models for an accurate understanding of soil nutrient quality information.

Calculation results from our soil nutrient evaluation investigation indicated that most of the ML techniques investigated were adequate in determining outputs. The GRNN and KNNSVM models yielded the highest overall evaluation accuracy rates of four ANN and six SVM models. According to the evaluation accuracy, the KNNSVM model is better than the GRNN model. Numerous studies have shown that SVM models can achieve good results in a variety of agricultural tasks (Camps-Valls et al. 2003; Karimi et al. 2006; Rumpf et al. 2010). In this study, the SVM models also exhibited good performance in determining outputs. Because the *k*-NN in the kernel line space was found directly in the KNNSVM model, the nonlinear phenomenon of a distance measure in many problems was avoided. After that, the relationship between neighbor and unlabeled samples was closer, and the classification accuracy improved overall. An SVM model was used for each cluster center through which clustering training samples were

established in the PSVM model, and then the unlabeled samples were classified using these SVMs. A balance between positive and negative samples was achieved by clustering. Part of the training samples were involved in the construction of the classifier, which was selected in the LMSVM model using the constraints of the relevant conditions, with locality. The FSVM took less time to obtain a clustering center than PSVM model, but the accuracy of the PSVM model was higher than LSVM model. For the LSVM model, the improvement in accuracy was not obvious with an increase in k value, and it was occasionally unstable. The training samples involved in building the SVMs were determined by SVM-KNN model, leading into the k-NN model (Shu 2015). As a result, the KNNSVM model combines the features of k-NN and SVM, having the advantage of high prediction accuracy. In this study, we determined that the KNNSVM model was the best model in estimating soil nutrition. However, performance of SVM models rely on input data to extract support vectors. The number of support vectors in the SVM models also increases with an increase in training sample numbers. When the number of training samples is large, the support vector becomes more complex.

The 10 models investigated in this study have their respective advantages and disadvantages. There are several reasons why we determined that the KNNSVM model is the best model among the 10 ML models investigated. The LMSVM model screens the training samples that participate in SVM classifier learning through an operator. It selects part of the training samples to participate in classifier construction, applying the constraints of relevant conditions. Thus, the learned SVM classifier has limitations. The SVM-KNN model resolves the deficiency of the LMSVM model by introducing the k-NN algorithm to determine the training samples involved in building the SVM. The KNNSVM algorithm calculates the distance between the samples in the nuclear feature space and looks for the neighbors of unlabeled samples, which avoids the instability caused by nonlinear problems among different distributions. The k-NN algorithm used in the LSVM model is essentially a weighting algorithm, having the disadvantage in the large amount of calculation required. The clustering achieved by the PSVM algorithm achieves a balance between positive and negative samples, a local SVM around the local cluster can be trained by its clustering. The number of classifiers is small, and the classification accuracy is not very high. The FSVM algorithm is proposed based on the KNNSVM algorithm. The FSVM model is superior to the PSVM model when using methods to resolve cluster center points. Given that it employs a strategy that reduces the number of local SVMs rather than directly resolving unlabeled samples for the k-NN training SVM centered on it, the local SVM is established with the k-NN of its nearest C-center. Thus, when unmarked samples are classified, the classification accuracy is worse than the KNNSVM model, but the amount of calculation is correspondingly lower. Evaluation accuracy of the selected data by the GRNN model was higher compared to the original data. The GRNN model should be used when the prediction of highly accurate results are required while avoiding the situation where back-propagation predicts the same database, lengthy algorithms, and instable network forecast results. In this study, 10 different computer algorithms were used to assess the soil nutrient content in the selected study areas. Among these, the number of samples used in the partial determination of k-NN gradually decreased with an increase in k values, while that of LSVM constantly increased. This showed that the uncertainty in unlabeled sample categories increases with an increase in the number of selected neighbors. The higher the k value, the more partial a SVM must be established (Gunn 1998; Rumpf et al. 2010; Shu 2015; Hao 2016).

It should be noted that it has many limitations in this study. For example, model simulation samples were not sufficiently large enough. Because the species is endangered, making specimen quantity is inadequate to meet the requirements of sampling. Moreover, uncertainty derives from factors related to field data acquisition, where environmental factors surrounding the soil are a significant factor in themselves, such as how soil temperature, humidity, sunlight, precipitation, as well as other climatic factors, affect the formation and availability of soil nutrients. Therefore, the next step in our investigation will be to increase the number of samples and add climate change factors in our investigation of soil nutrient quality.

Conclusions

The result of this study shows that ML models are well suited for soil nutrient evaluation. The KNNSVM model can be used effectively to soil nutrient evaluation by using appropriate model variables, and the GRNN model is also a good choice albeit less suitable than the former model due to its low RMSE values. Therefore, the KNNSVM model can be used to determine outputs among the 10 ML models investigated. We have determined that our model has significant potential in getting outputs, and it can be considered as an alternative tool in determining the soil nutrient condition of rare and endangered tree species on regional or global scales. These models can be applied to many applications, such as providing support decision information to forest managers or conducting conservation strategies for large-scale rare and endangered tree in natural forest. The proposed method can improve the accuracy domain of the current multiple linear regression model in this

study. Invisible data from proven high-precision machine learning models may improve the usefulness and accuracy of decision-making to provide information to support agricultural stakeholders.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40663-020-00232-5.

Additional file 1: Table S1. Artificial neural network calculation results one. **Table S2.** Artificial neural network calculation results two. **Table S3.** Artificial neural network calculation results three.

Abbreviations

k-NN: *K*-nearest neighbors; KNNSVM: *K*-nearest neighbors -support vector machine; VNIR: Visible and near-infrared; MLR: Multiple linear regression; PLS regression: Partial least squares regression; ML: Machine learning; ANN: Artificial neural networks; SVR: Support vector regression; FIS: Fuzzy inference systems; C: Carbon; N: Nitrogen; P: Phosphorus; NO_3^- N: Nitrate nitrogen; *D. pectinatum*: *Dacrydium pectinatum*; SOM: Soil organic matter; K: Potassium; Ca: Calcium; Mg: Magnesium; S: Sulfur; Cu: Copper; Zn: Zinc; Fe: Iron; Mn: Manganese; MSE: Mean squared error; ME: Mean error; MSPE: Mean square prediction error; RMSE: Root-mean-squared error; ROC: Receiver operating characteristic; FPR: False positive rate; TPR: True positive rate; AUC: Area under the curve

Acknowledgments

The authors of this study would like to thank the Hainan Bawangling Natural Nature Reserve of Hainan Province who provided the test sites and experimental materials, and the work was financially supported by the Fundamental Research Funds for the Central Non-profit Research Institution of CAF (CAFBB2017ZB004). CW would also like to thank the China Scholarship Council (CSC) for offering a scholarship at the University of Quebec at Montreal (UQAM). CP acknowledges the funding provided by the National Science and Engineering Research Council of Canada (NSERC) Discover Grant.

Authors' contributions

CW and YC conceived and designed the experiments. CW and XH performed the experiments. ZL supplied the methods references. CP edited and revised the manuscript. The author(s) read and approved the final manuscript.

Funding

The work was financially supported by the Fundamental Research Funds for the Central Non-profit Research Institution of CAF (CAFBB2017ZB004).

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Tree Breeding and Cultivation of the State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China. [2]Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China. [3]Hainan Bawangling National Natural Reserve, Changjiang 572722, Hainan, China. [4]Department of Biological Science, Institute of Environment Sciences, University of Quebec at Montreal, Montreal, QC, Canada.

Received: 17 April 2019 Accepted: 23 March 2020
Published online: 05 May 2020

References

Akbarzadeh S, Paap A, Ahderom S, Apopei B, Alameh K (2018) Plant discrimination by support vector machine classifier based on spectral reflectance. Comput Electron Agr 148:250–258

Ash J (1986) Growth rings, age and taxonomy of *Dacrydium* (Podocarpaceae) in Fiji. Aust J Bot 34(2):197–205

Bassaco MVM, Motta ACV, Pauletti V, Prior SA, Nisgoski S, Ferreira CF (2018) Nitrogen, phosphorus, and potassium requirements for *Eucalyptus urograndis* plantations in southern Brazil. New Forest 49(5):681–697. https://doi.org/10.1007/s11056-018-9658-0

Bibby J, Toutenburg H (1977) Prediction and improved estimation in linear models. Wiley, New York. https://doi.org/10.1002/bimj.197800029

Brailovsky VL, Barzilay O, Shahave R (1999) On global, local, mixed and neighborhood kernels for support vector machines. Pattern Recogn Lett 20(11–13):1183–1190

Camenzind T, Hättenschwiler S, Treseder KK, Lehmann A, Rillig MC (2018) Nutrient limitation of soil microbial processes in tropical forests. Ecol Monogr 88(1):4–21. https://doi.org/10.1002/ecm.1279

Camps-Valls G, Gómez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martín-Guerrero JD, Moreno J (2003) Support vector machines for crop classification using hyperspectral data. In: Perales FJ, Campilho AJC, de la Blanca NP, Sanfeliu A (eds) Pattern recognition and image analysis. IbPRIA 2003. Lecture notes in computer science, vol 2652. Springer, Berlin, Heidelberg, pp 134–141

Cao XP, Zou Y, Yu JL, Yang SX, Chen H, Huang YC, Shen AM, Shen JQ (2017) Resources, protection and utilization of rare and endangered tree species in northeastern Jiangxi Province. South China Forest Sci 45(43–47):52 (in Chinese)

Chagnon P-L, Brown C, Stotz GC, Cahill JF Jr (2018) Soil biotic quality lacks spatial structure and is positively associated with fertility in a northern grassland. J Ecol 106(1):195–206. https://doi.org/10.1111/1365-2745.12844

Chen YK, Yang Q, Mo YN, Yang XB, Li DH, Hong XJ (2014) Study on the niche of nationally protected plants in Bawangling, Hainan Island. Chin J Plant Ecol 38:576–584 (in Chinese)

Cheng H, Tan PN, Jin R (2010) Efficient algorithm for localized support vector machine. IEEE T Knowl Data En 22(4):537–549

Cheng Q (2005) Research on the ANN design algorithm. Microcomputer Development 15(12):61–62, 65 (in Chinese)

Comizzoli P (2015) Biobanking efforts and new advances in male fertility preservation for rare and endangered species. Asian J Androl 17(4):640–645

Comizzoli P, Holt WV (2014) Recent advances and prospects in germplasm preservation of rare and endangered species. In: Holt W, Brown J, Comizzoli P (eds) Reproductive sciences in animal conservation . Advances in experimental medicine and biology, vol 753. Springer, New York, NY, pp 331–356

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, Cambridge

Deng JQ, Chen XM, Wang R, Nan JK, Du ZJ (2017) LS-SVM data mining analysis: how does biochar influence soil net nitrogen mineralization in the field? J Soils Sediments 17:827–840

Dou XM, Yang YG (2017) Modeling and predicting carbon and water fluxes using data-driven techniques in a forest ecosystem. Forests 8:498. https://doi.org/10.3390/f8120498

Du ZX (2016) The study and the evaluation of the nutrient element content in the Liao river estuary wetland soil under the papermaking wastewater irrigation. Master's degree thesis. Shenyang: Shenyang Agricultural University (in Chinese)

Ebrahimi M, Sinegani AAS, Sarikhani MR, Mohammadi SA (2017) Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses. Comput Electron Agr 140:409–421. https://doi.org/10.1016/j.compag.2017.06.019

Farjon A, Filer D (2013) An atlas of the world's conifers: an analysis of their distribution, biogeography, diversity and conservation status. Brill Academic Publication, Boston

Fu LG (1992) China plant red data book and endangered plants, vol 1. Science Press, Beijing (in Chinese)

Gerloff GC, Krombholz PH (1966) Tissue analysis as a measure of nutrient availability for the growth of angiosperm aquatic plants. Limnol Oceanogr 11(4):529–537. https://doi.org/10.4319/lo.1966.11.4.0529

Ghahramani Z (2015) Probabilistic machine learning and artificial intelligence. Nature 521:452–459

Giovanis DG, Papaioannou I, Straub D, Papadopoulos V (2017) Bayesian updating with subset simulation using artificial neural networks. Comput Method Appl M 319:124–145

Grove S, Parker IM, Haubensak KA (2017) Do impacts of an invasive nitrogen-fixing shrub on Douglas-fir and its ectomycorrhizal mutualism change over time following invasion? J Ecol 105(6):1687–1697. https://doi.org/10.1111/1365-2745.12764

Gunn SR (1998) Support vector machines for classification and regression. Department of Electronics and Computer Science of University of Southampton, Southampton, pp 1–28

Guo DD, Juan JX, Chang LY, Zhang JJ, Huang DF (2017) Discrimination of plant root zone water status in greenhouse production based on phenotyping and machine learning techniques. Sci Rep 7:8303. https://doi.org/10.1038/s41598-017-08235-z

Hao QB (2016) Research and application of local support vector machine in classification. Master's degree thesis. Taian: Shandong Agricultural University (in Chinese)

Idowu OJ, van Es HM, Abawi GS, Wolfe DW, Ball JI, Gugino BK, Moebius BN, Schindelbeck RR, Bilgili AV (2008) Farmer-oriented assessment of soil quality using field, laboratory, and VNIR spectroscopy methods. Plant Soil 307:243–253

Jha SK, Ahmad Z (2018) Soil microbial dynamics prediction using machine learning regression methods. Comput Electron Agr 147:158–165

Karimi Y, Prasher SO, Patel RM, Kim SH (2006) Application of support vector machine technology for weed and nitrogen stress detection in corn. Comput Electron Agr 51(1–2):99–109

Karlen DL, Andrews SS, Doran JW (2001) Soil quality: current concepts and applications. Adv Agron 74:1–40

Kawamura K, Tsujimoto Y, Rabenarivo M, Asai H, Andriamananjara A, Rakotoson T (2017) Vis-NIR spectroscopy and PLS regression with waveband selection for estimating the total C and N of paddy soils in Madagascar. Remote Sens 9(10):1081. https://doi.org/10.3390/rs9101081

Kim M, Gilley JE (2008) Artificial neural network estimation of soil erosion and nutrient concentrations in runoff from land application areas. Comput Electron Agr 64(2):268–275

Li H, Leng W, Zhou YB, Chen FD, Xiu ZL, Yang DZ (2014) Evaluation models for soil nutrient based on support vector machine and artificial neural networks. Sci World J 478569. https://doi.org/10.1155/2014/478569

Li YF, Liang S, Zhao YY, Li WB, Wang YJ (2017) Machine learning for the prediction of *L. chinensis* carbon, nitrogen and phosphorus contents and understanding of mechanisms underlying grassland degradation. J Environ Manag 192:116–123

Lian JS, Yu SX (2011) Floristic characters of the formation *Dacrydium pierrei/Syzygium araiocladum* in tropical montane rain forest in Bawangling nature reserve, Hainan Island. J Trop Subtrop Bot 9:101–107 (in Chinese)

Marcos MS, Bertiller MB, Cisneros HS, Olivera NL (2016) Nitrification and ammonia-oxidizing bacteria shift in response to soil moisture and plant litter quality in arid soils from the Patagonian Monte. Pedobiologia 59(1–2):1–10

Moges MA, Schmitter P, Tilahun SA, Langan S, Dagnew DC, Akale AT, Steenhuis TS (2017) Suitability of watershed models to predict distributed hydrologic response in the Awramba watershed in Lake Tana Basin. Land Degrad Develop 28(4):1386–1397

Murphy CJ, Baggs EM, Morley N, Wall DP, Paterson E (2017) Nitrogen availability alters rhizosphere processes mediating soil organic matter mineralisation. Plant Soil. https://doi.org/10.1007/s11104-017-3275-0

Myers PD, Scirica BM, Stultz CM (2017) Machine learning improves risk stratification after acute coronary syndrome. Sci Rep 7:12692. https://doi.org/10.1038/s41598-017-12951-x

Olego MA, Visconti F, Quiroga MJ, de Paz JM, Garzón-Jimeno E (2016) Assessing the effects of soil liming with dolomitic limestone and sugar foam on soil acidity, leaf nutrient contents, grape yield and must quality in a Mediterranean vineyard. Span J Agric Res 14(2):e1102. https://doi.org/10.5424/sjar/2016142-8406

Pingree MRA, DeLuca TH (2018) The influence of fire history on soil nutrients and vegetation cover in mixed-severity fire regime forests of the eastern Olympic peninsula, Washington, USA. Forest Ecol Manag 422(15):95–107

Qi HJ, Paz-Kagan T, Karnieli A, Jin X, Li SW (2018) Evaluating calibration methods for predicting soil available nutrients using hyperspectral VNIR data. Soil Till Res 175:267–275

Riccioli F, Marone E, Boncinelli F, Tattoni C, Rocchini D, Fratini R (2019) The recreational value of forests under different management systems. New Forest 50:345–360. https://doi.org/10.1007/s11056-018-9663-3

Roudier P, Malone BP, Hedley CB, Minasny B, McBratney AB (2017) Comparison of regression methods for spatial downscaling of soil organic carbon stocks maps. Comput Electron Agr 142(A):91–100

Rumpf T, Mahlein AK, Steiner U, Oerke EC, Dehne HW, Plümer L (2010) Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. Comput Electron Agr 74(1):91–99

Segata N, Blanzieri E (2010) Fast and scalable local kernel machines. J Mach Learn Res 11:1883–1926

Shine P, Murphy MD, Upton J, Scully T (2018) Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms. Comput Electron Agr 150:74–87

Shu ZY (2015) The classification of high resolution remote sensing images based on local support vector machine. Doctoral dissertation. University of Geosciences, China (in Chinese)

Sirsat MS, Cernadasa E, Fernández-Delgadoa M, Khan R (2017) Classification of agricultural soil parameters in India. Comput Electron Agr 135(1):269–279

Sousa-Silva R, Alves P, Honrado J, Lomba A (2014) Improving the assessment and reporting on rare and endangered species through species distribution models. Glob Ecol Conserv 2:226–237

Sun K, Wang ZJ, Tu K, Wang SJ, Pan LQ (2016) Recognition of mould colony on unhulled paddy based on computer vision using conventional machine-learning and deep learning techniques. Sci Rep 6:37994. https://doi.org/10.1038/srep37994

Vacca A, Aru F, Ollesch G (2017) Short-term impact of coppice management on soil in a *Quercus ilex* l. stand of Sardinia. Land Degrad Dev 28(2):553–565

Wang JF, Bao ST, Chen SS, Wang YF (2008) FT-NIR Spectroscopy technique based analysis and prediction on soil nutrient content of Lychee orchard: a case study in Zhongluotan of Guangzhou, South China. In: Proc. SPIE 7145, Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Monitoring and Assessment of Natural Resources and Environments, 71451D (3 November 2008). https://doi.org/10.1117/12.813028

Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic 52:394–403

Xu EQ, Zhang HQ, Li MX (2015) Object-based mapping of karst rocky desertification using a support vector machine. Land Degrad Dev 26(2):158–167

Zhang DY, Zhang W, Huang W, Hong ZM, Meng LK (2017) Upscaling of surface soil moisture using a deep learning model with VIIRS RDR. ISPRS Int J Geo-Inf 6(5):130. https://doi.org/10.3390/ijgi6050130

Zhao ZY, Chow TL, Rees HW, Yang Q, Xing ZS, Meng FR (2009) Predict soil texture distributions using an artificial neural network model. Comput Electron Agr 65(1):36–48