Article

Remote Sensing Classification and Mapping of Forest Dominant Tree Species in the Three Gorges Reservoir Area of China Based on Sample Migration and Machine Learning

Wenbo Zhang [1,2,3], **Xiaohuang Liu** [1,3,*], **Bin Xu** [4], **Jiufen Liu** [1,2,3], **Hongyu Li** [1,2,3], **Xiaofeng Zhao** [1,2,3], **Xinping Luo** [1,3], **Ran Wang** [1,2,3], **Liyuan Xing** [1,3], **Chao Wang** [1,3] and **Honghui Zhao** [1,3]

1   Comprehensive Survey Command Center for Natural Resources, China Geological Survey, Beijing 100055, China; zhangwenbo23@mail.cgs.gov.cn (W.Z.); liujiufen@mail.cgs.gov.cn (J.L.); lihongyu@mail.cgs.gov.cn (H.L.); zhaoxf2011@163.com (X.Z.); luoxinping@mail.cgs.gov.cn (X.L.); wangran@mail.cgs.gov.cn (R.W.); xingliyuan@mail.cgs.gov.cn (L.X.); wangchao01@mail.cgs.gov.cn (C.W.); zhaohonghui@mail.cgs.gov.cn (H.Z.)
2   School of Earth Science and Resources, China University of Geosciences (Beijing), Beijing 100083, China
3   Key Laboratory of Coupling Process and Effect of Natural Resources Elements, Beijing 100055, China
4   Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China; xubin@caas.cn
*   Correspondence: liuxiaohuang@mail.cgs.gov.cn

Abstract: The distribution of forest-dominant tree species is crucial for ecosystem assessment. Remote sensing monitoring requires annual ground sample data, but consistent field surveys are challenging. This study addresses this by combining sample migration learning and machine learning for multi-year tree species classification in the Three Gorges Reservoir area in China. Using the continuous change detection and classification (CCDC) algorithm, sample data from 2023 were successfully migrated to 2018–2022, achieving high migration accuracy (R^2 = 0.8303, RMSE = 4.64). Based on migrated samples, random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB) algorithms classified forest tree species with overall accuracies above 70% and Kappa coefficients above 0.6. XGB. They outperformed other algorithms, with classification accuracy of over 80% and Kappa above 0.75 in almost all years. The final map indicates stable distribution from 2018 to 2023, with eucalyptus covering over 40% of the forest area, followed by horsetail pine, fir, cypress, and wetland pine.

Keywords: forest ecosystem; forest dominant tree species; machine learning; migration learning; change detection

1. Introduction

Information on the spatial distribution of forest dominant species is an important basis for sustainable management of forest resources, carbon stock monitoring, and species diversity assessment [1–4], and multi-year information on forest tree species is considered to be important in formulating forestry management strategies and promoting the development of forest economic benefits [5]. Traditional periodic forest inventory tasks generally use manual recording means in estimating the distribution of forest species, which often suffers from the following problems: high labor costs, poor accessibility, and long task aggregation cycles [6,7]. This also leads to the inability of traditional means to provide continuous multi-year tree species distribution information in the context of large spatial scales [8].

Remote sensing data have received widespread attention in forest monitoring tasks due to their large observation range and fast data update, and are considered to be an important data source for mapping forest stand attributes (tree species distribution, etc.) [9,10]. However, detailed stand tree species information is usually utilized with very high resolution (VHR) data such as WorldView-2 [11,12]. Limited by the cost of access, most

VHR-based studies tend to be confined to small areas, such as Waster et al. [11], who used WorldView-2 data for tree species classification work within a study area of 60 km^2. Tree species classification studies for large geographical scales with complex topographic conditions and locational factors rarely use VHR data [9,13]. Instead, existing large-scale remote sensing data products usually focus only on broad parameter information of forests, such as forest cover, forest type (plantation/natural forest, coniferous/broadleaf forest), and forest carbon potential [14–16].

The Landsat series of moderate-resolution satellites are widely used for large-area forest monitoring missions due to their free and open-source characteristics [17]. However, its temporal resolution of 16 days and spatial resolution of 30 m have exposed some problems: (1) lower temporal resolution results in a limited number of cloud-free and seamless Landsat remote sensing imagery acquisitions, which does not allow for the complete capture of phenological information [18]. (2) Due to the fact that spectral separability and stand structure are highly correlated, differences in stand structure between tree species have been shown to serve as potential features for tree species classification tasks, thus helping to improve the classification accuracy of tree species [19]. However, the spatial resolution of 30 m performs poorly in identifying structural differences in forest stands, resulting in lower accuracy of tree species classification [20]. Fortunately, the Sentinel-2 satellites launched in 2015 and 2017 brought significant advances in spatial and temporal resolution from medium-resolution satellites. The 5-day revisit period means that more observations are achieved in the same time span, allowing for the complete capture of rapidly changing phenological information [21]. Grabska et al. [22] found in their task of mapping tree species in mountainous areas that images acquired at the beginning and end of the climatic period (spring and autumn) can distinguish well among different forest tree species. The 10 m spatial resolution of Sentinel-2 data can capture more detailed information of tree species, and the correlation between texture features calculated based on the 10 m resolution and stand structure information is stronger than that calculated based on Landsat data, which can help to distinguish between tree species [23].

In performing the task of mapping multi-year tree species distribution information, a challenge arises that cannot be ignored. Usually, the sample data used for the tree species classification task are mainly derived from (1) National Forest Inventory (NFI) data and (2) in situ measurements of forest plots [24,25]. The sample data obtained based on the above methods are often limited by the acquisition cost, which makes it difficult to update the data in a timely manner. Additionally, the acquisition of multi-year tree species distribution information faces two problems: (1) is it feasible to use the samples of a single year to classify the images of other years? (2) How can we accurately map single-year tree species sample data to multiple years? In other words, how can we achieve interannual migration of single-year samples? To address the first question, existing studies have demonstrated the feasibility of supervised classification of images in the current time phase based on past training samples [26,27]. For example, Zhang et al. [26] used samples obtained from previous land use data products to classify images in the current time phase and achieved high classification accuracy. However, a problem was also exposed: the quality of the sample data may be affected by previous classification errors. Therefore, it is the second question that we want to answer., In a multi-year wetland land cover classification task to ensure the quality of the samples during the migration process, Fekri et al. [27] obtained a stabilized sample by calculating the similarity of the three vegetation indices for the reference year–target year (with no training samples) and deriving the optimal thresholds for the unchanged samples, which were used in 2018, 2019, and 2021 target years. The proposed migrated sample approach yielded more than 95% accuracy. However, there was a huge difference in the tree species classification task compared to the land cover classification task, and the spectral differences among tree species were small compared to the spectral differences among different land cover types, so the method based on spectral similarity thresholding may need to be based on a large number of vegetation indices combined with spectral band features to obtain convincing sample migration

results. This, in turn, represents the need to cope with large computational volumes and the processing of high-dimensional and complex data. The online change detection algorithms (CCDC algorithm, Landtrendr algorithm) cope well with these problems, and this class of algorithms has been well integrated into the Google Earth Engine (GEE), which does not require complex data input and achieves accurate identification of forest disturbance situations by setting a series of disturbance discrimination rules. For example, Yang et al. [21] achieved an R^2 of 0.79 using the CCDC change detection algorithm for the forest disturbance identification task. The research strongly confirms the potential of this class of transform detection algorithms in identifying perturbed image elements and stable image elements. Additionally, it is not yet known whether the CCDC algorithm can perform the task of migrating stable forest tree samples in time series.

In this study, we used multi-temporal remote sensing data for each year from 2018–2023 as the data source, invariant samples obtained from each year using the CCDC change detection algorithm as the reference data, and the random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB) algorithms as the classification algorithms. The aim is to map the forest-dominant tree species in the Three Gorges Reservoir Area from 2018 to 2023. In general, this study is committed to solving the following problems:

(1) Is it possible to obtain high-precision tree species classification results using samples after completing inter-annual migration based on the CCDC change detection algorithm?
(2) Which machine learning algorithm performs best in the task of classifying forest-dominant tree species in the Three Gorges Reservoir area?

2. Study Area and Data

2.1. Study Area

Our study area is located in southwestern China (e.g., Figure 1a) at the intersection of the Sichuan Basin and the plains of the middle and lower reaches of the Yangtze River. Its official name is the Three Gorges Reservoir area (TGRA), which mainly includes 20 counties (districts) that are inundated by the Three Gorges Project of the Yangtze River and have the task of migrant resettlement. The geographic location is 106°20′–110°30′E, 29°00′–31°50′N, from Yiling District in the east to Jiangjin District in the west, with an area of more than 60,000 km^2 and a population of more than approximately 23 million in the study area. The climate of the study area is a humid subtropical monsoon climate, with an average annual precipitation of more than 1000 mm and an average annual temperature of around 17 degrees Celsius. The study area is more than 74% mountainous, 22% hilly, and only 4% plain, with the overall topography ranging from high in the east to low in the west, with altitudes ranging from 12 to 2994 m above sea level (Figure 1b). The forest cover of the study area is more than 60% (Figure 1c), and the forests are mainly distributed in the eastern mountains, the western karst mountains, and the southern low hills, with the main forest types being evergreen forests. Due to decades of over-exploitation and logging, the proportion of natural forests in the area is about 4 percent, and most of the natural forests are secondary forests. According to forestry statistics, information on forest species in the study area is relatively simple. Five tree species, cypress, horsetail pine, wetland pine, fir, and eucalyptus, account for more than 90 per cent of the total forest area and total stock in the study area.
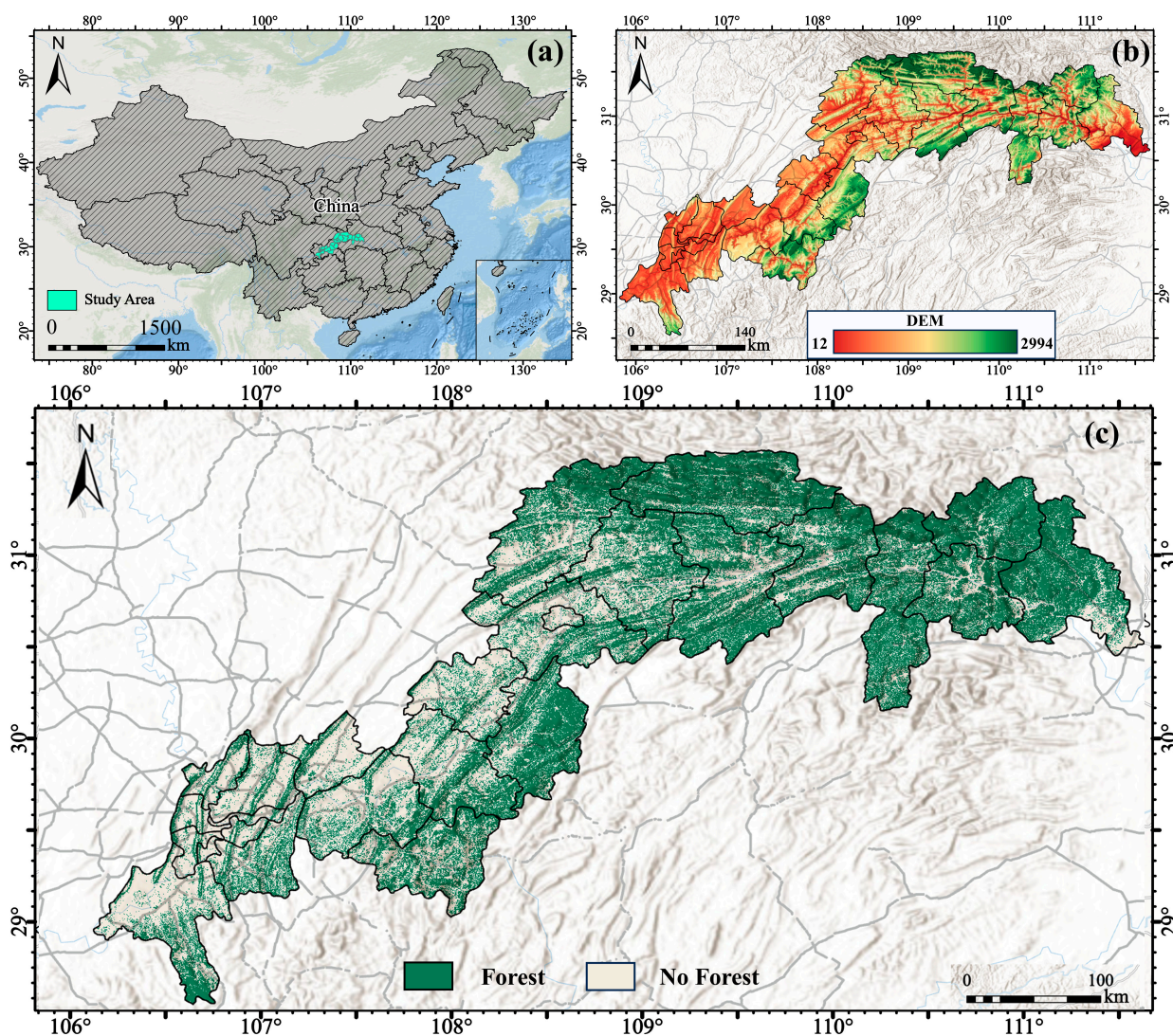
Figure 1. Overview map of the study area. (**a**) represents the location of the study area in China; (**b**) represents the DEM data of the study area; (**c**) represents the forest cover in the study area.

2.2. Data

2.2.1. Sentinel-2 Data

In this study, all Sentinel-2 image tiles with less than 20% cloud cover in the study area from 2018–2023 were collected using the GEE platform. The SR dataset used is the L2A-level surface reflectance data product that has undergone an atmospheric correction operation. Additionally, cloud/cloud shading was masked for each image by quality assessment bands (except for 2018, which used the MSK_CLASSI_SNOW_ICE band, and the rest of the years, which used the QA60 band). Subsequently, in order to ensure the consistency of each band in the index calculation, this study used bilinear interpolation to resample the band with a spatial resolution of 20 m to 10 m. Due to the influence of conditions such as cloud cover and orbit distribution on the image quality, it becomes exceptionally difficult to obtain cloud-free and seamless remote sensing data for the whole study area. To cope with the above situation, this study used the median synthesis method to generate a cloud-free image for each season. In the end, a total of 24 cloud-free and seamless high-quality Sentinel-2 images covering the entire study area were obtained.

The spectral bands that were used for the classification task in this study include B2–B4 (RGB), B5–B7 (red edge), B8 (near-infrared), and B11–B12 (shortwave-infrared). Based on the above data, NDVI (normalized difference vegetation index), NDWI (nor-

malized difference water index), SAVI (soil-adjusted vegetation index), NIRV (nearby vegetation index), and REIP (red-edge inflection point index) were computed for each image in this study which has been proved to be the reliable indices to improve the accuracy of tree species classification in previous studies. In the subsequent process of calculating the texture indices, the first principal component of each image was extracted using the principal component analysis (PCA) method, and the greyscale covariance matrix of the first principal component was calculated. The eight texture features, energy, entropy, correlation, inverse difference moment, inertia, cluster shade, cluster prominence, and correlation, were finally obtained. Subsequently, we stacked the spectral bands, vegetation indices, and texture features corresponding to each year to construct a multi-temporal remote sensing image dataset. The stacked images will be used as a remote sensing data source for tree species classification. The computation of both vegetation indices and texture features was conducted using the GEE platform.

2.2.2. Landsat Data

Because the change detection algorithms used in the subsequent interannual migration of the samples were based on data acquired over the entire Landsat observation cycle, we collected all available Landsat imagery in the GEE platform for the study area for the period 1986–2023 (the earliest available Landsat data in the GEE platform is 1986). These images were obtained from the Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager (OLI) sensors. In order to enhance the image quality, we first implemented a de-cloud/shadow operation for each image using the QA quality assessment band, followed by a composite operation on a monthly basis for all images. In addition, to enhance the accuracy of the CCDC algorithm in identifying changing samples, we also calculated NDVI and NBR, which tend to show higher values in healthy and lush forests and are more sensitive to forest disturbance and recovery. The resulting long-time series remote sensing data and spectral indices will both serve as valuable data inputs for subsequent change detection algorithms. The formula for the spectral index is shown in Table 1.

Table 1. Calculation formula for the spectral index as a data input to the change detection algorithm.

Spectral Indices	Calculation Formula
NDVI	$(NIR - RED)/(NIR + RED)$
NBR	$(NIR - SWIR)/(NIR + SWIR)$ [1]

[1] Due to nomenclature differences in remote sensing imagery collected by different Landsat sensors, SWIR in the equation represents the wavelength range of 2.08–2.35 microns.

2.2.3. Tree Species Sample Data

The tree species sample data used in this study were obtained by (1) the 2023 National Forestry Inventory (NFI) data, (2) field measurement data, and (3) UAV hyperspectral data. For the forestry inventory data in vector data format, the sample patches with the field value 'pure forest' were first selected, and patches with an area of less than 900 m^2 were deleted. Forest mask data were obtained after reclassification based on the existing 30 m resolution land cover dataset [28], and sample patches containing non-forest image elements inside were deleted. In addition, the 'planted area' and 'volume' fields were counted, and the top five tree species in terms of planted area and volume were filtered for classification (this was due to the fact that the top five species had more than 90% of the overall planted area and volume). The filtered samples were imported into 'Google Earth pro' software for visual interpretation, and sample patches with non-forested areas were further removed. The final 3367 high-quality sample plots were characterized by the following features: pure forest stands, no non-forested areas within the plots, and plots with an area of more than 900 m^2. In the field measurements, the target plots were required to have more than 80% monoculture trees, and all trees within the plots had to be greater than 2 m in height and more than 5 cm in diameter at breast height (DBH). At

the same time, the trees in the plots were required to be free of recent felling, pests, and diseases. Subsequently, the type of tree species was recorded with the assistance of the local forestry department. The GPS location of the plots was recorded using 'Ovital Map 10.0.5' software. The GPS information was converted into vector point data in ArcGIS pro 3.0 software, and the sample points were set to have the same coordinate information as the remote sensing image. Additionally, a total of 1324 real samples were obtained. For forest areas with poor accessibility, this study used a hyperspectral mapping UAV with a spatial resolution of 0.03 m to complete the measurement task. Each hyperspectral image had a size of 20 × 20 m, and under the guidance of the staff, images containing mixed forests were excluded, and the tree species types in the pure stand images were recorded. Subsequently, after setting a uniform coordinate system for the hyperspectral images in ArcGIS pro-3.0, vector points were generated at the center of the images, and the vector points were assigned with tree species information. Finally, 677 sample data based on hyperspectral images were obtained. Based on the above operations, a total of 5368 high-quality samples were collected in this study, and the tree species types included cypress, horsetail pine, wetland pine, fir, eucalyptus, camphor, oak, maple, and quebracho. Among them, cypress, horsetail pine, wetland pine, fir, and eucalyptus were the taxonomic objects of this study. The number of samples of each tree species is shown in Table 2, and the distribution of tree species in the study area is shown in Figure 2.

Table 2. Number of samples for each tree species.

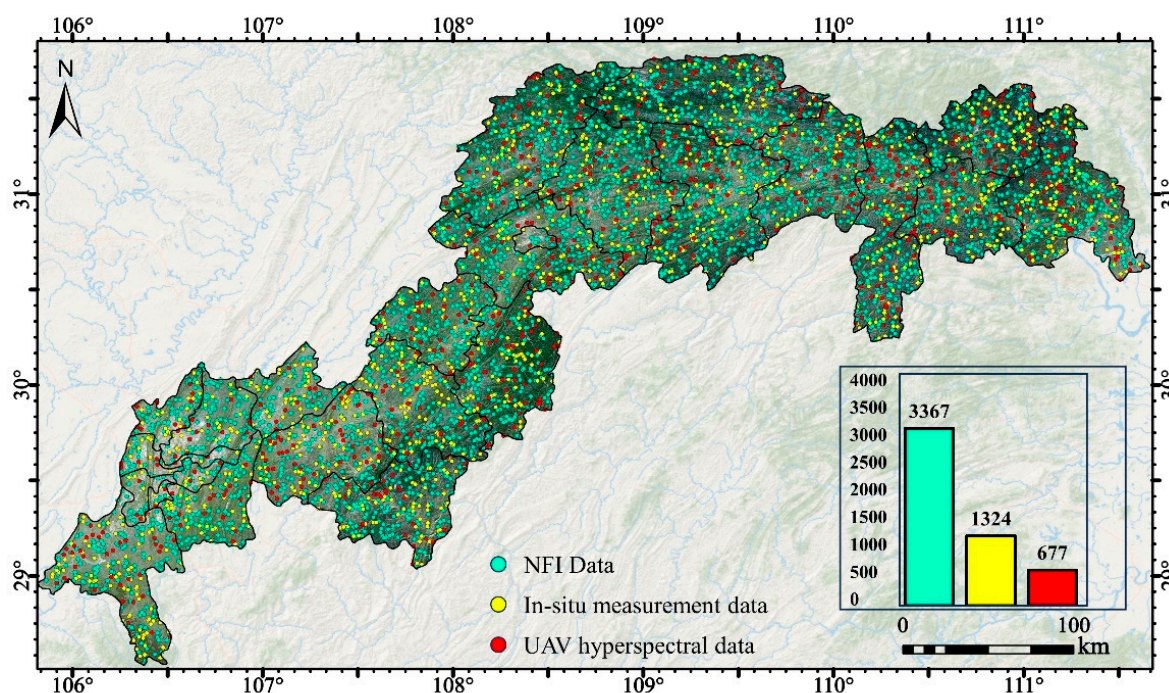| Type	Name	Number
dominant tree species	cypress	1037
	horsetail pine	956
	wetland pine	862
	fir	1042
	eucalyptus	1349
minor species	camphor, quebracho, maple, oak, etc.	122

Figure 2. Distribution of sample data for tree species in the study area.

3. Methods

The workflow for conducting the task of mapping dominant tree species in the study area consisted of the following steps (Figure 3): (1) inter-annual migration of samples, (2) computation of taxonomic features, and (3) training and accuracy assessment of classification models.
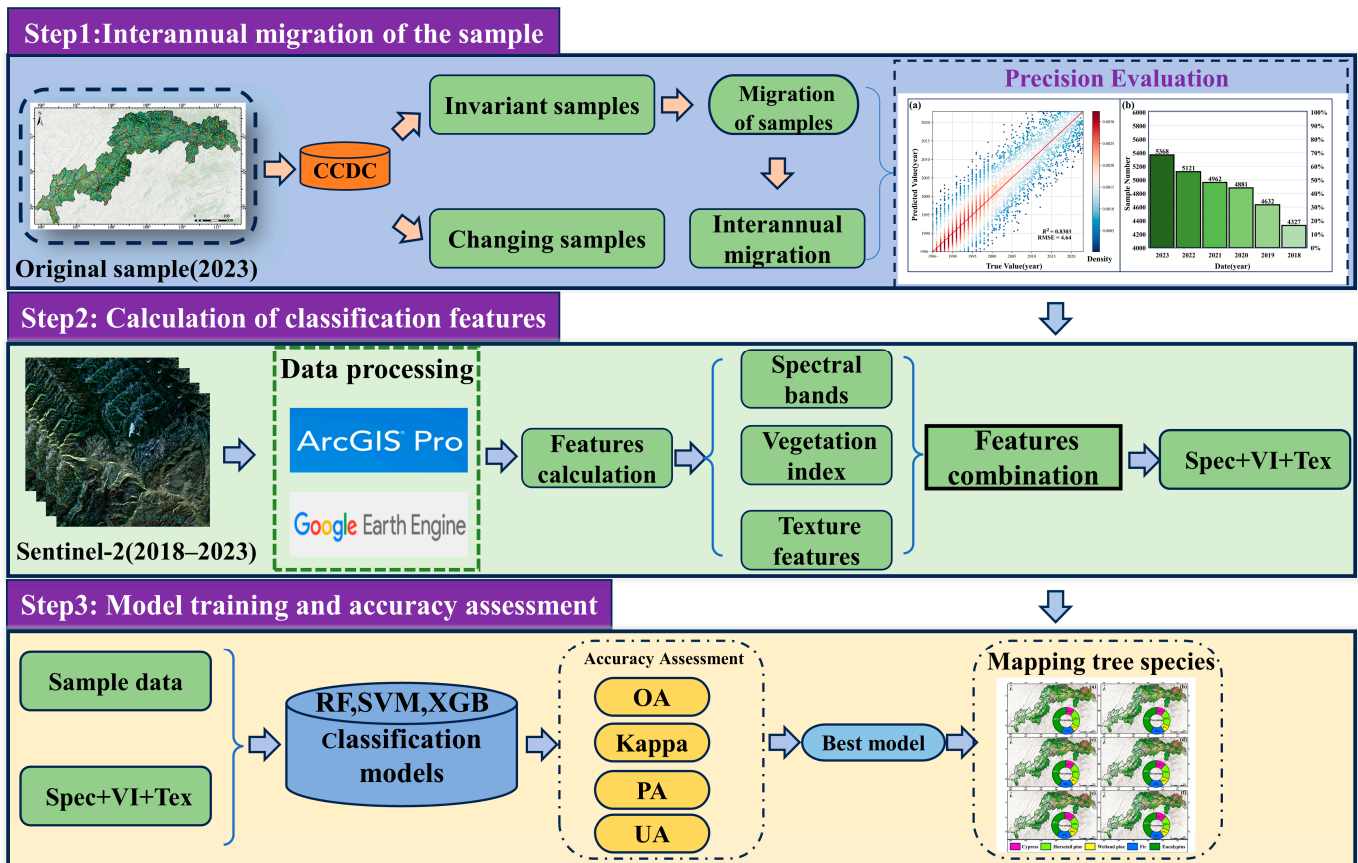

Figure 3. Workflow for mapping forest dominant tree species in the Three Gorges Reservoir area, 2018–2023. CCDC: the continuous change detection and classification algorithm; Spec: spectral band; VI: vegetation index; Tex: texture feature; RF: random forest algorithm; SVM: support vector machine algorithm; XGB: extreme gradient boosting algorithm; OA: overall accuracy; PA: producer accuracy; UA: user accuracy.

3.1. Sample Migration

Most of the limited inter-annual sample migration studies are based on spectral similarity or measuring the distance of spectral angles to achieve sample migration. However, these methods are often based on local, high-dimensional, and complex data processing, and the resulting high computational costs are difficult to accept. Fortunately, the continuous change detection and classification (CCDC) algorithm deployed in the GEE platform copes well with this problem. Its wide range of time series data can quickly establish the long time series curve of each pixel in the area and determine whether the pixel is perturbed or not by judging whether each data point in the time series curve is an outlier or not [29]. The CCDC algorithm utilizes the long time series NDVI and NBR as the modeling data, and based on the ordinary least squares (OLS) method, a linear fitting model is built for each pixel in the study area. Additionally, the difference between the observed and fitted values is subsequently calculated. When the difference between the two was greater than three times the root mean square error (RMSE) for six consecutive times within a sustained observation period, the pixel was recorded as a disturbed image element [30]. At the same time, breaks appear in the time series curve obtained from the linear fit. However,

NDVI and NBR may have different sensitivities to forest disturbance situations, which may cause the algorithms to diverge when determining the time of disturbance occurrence. To counteract this problem, we analyzed the 'chiSquareProbability' band in the change detection results, which represents the probability of each breakpoint occurrence [30]. In this study, we compared the probability of occurrence of breakpoints in the NDVI and NBR time series curves and recorded the occurrence time of breakpoints with a high probability of occurrence as the time of perturbation. However, the occurrence of a perturbed pixel within a sample patch does not mean that the sample patch needs to be withdrawn from the sample migration process. Since we have determined that all sample patches from 2023 are high-quality samples with pure stands and well-defined species types, when migrating sample data toward years 18–22 (target years), a sample patch can be migrated as long as the sample patch has not been disturbed prior to the target year. As an example, if a sample patch is disturbed in 2019, and after a year of restoration (hand transplanting/natural growth), the disturbed area grows back and no disturbance is detected to have occurred in 2020–2022. At this point, the sample patch is then ready for sample relocation in 2020–2022. Meanwhile, since the tree type of the 2018–2019 sample is not known, this sample patch cannot be sample relocated in 18–19. In short, a sample patch that has been disturbed during 2018–2022 can only be migrated backwards, not forwards.

At this point, an important issue comes to light: when the area is disturbed, it may not immediately revert to forest. At this time, the area will appear to be 'idle' for a period of time, and it is the time when it actually returns to the forest, which is an important criterion for determining whether the sample patch can be used for migration. Based on this problem, this study draws on the decision of Du et al. [31], who eliminated misdetection when determining the plantation time of plantation forests, identifying the year of restoration using the following criteria: (1) the increment of the time-series curve in which restoration occurs after a disturbance must be greater than 0.2; (2) the time-series curve duration must be more than one year; (3) the year of recovery must be the date corresponding to the first vertex of the time series curve judged to be in recovery. This decision-making method was shown to have a precise determination of the recovery of the disturbed area in the study of Yang et al. [21]. Based on the above method, the following situations may occur during the interannual migration of sample patches in the study area:

(1) For sample patches in which no disturbance occurrence was detected in the whole time-series curve (Figure 4a), this category of sample patches can be used as a classification sample for any of the years 2018–2023 (the recovery time of this category of sample patches was recorded as 1986–).

(2) For sample patches where the disturbance phenomenon occurred prior to 2018 (Figure 4b), sample patches in this category can also be used as classification samples for any of the years 2018–2023 (the recovery time for sample patches in Figure 4b is 1989).

(3) For sample patches where the perturbation phenomenon occurred between 2018–2022 (Figure 4c), sample patches of this class can only be used for sample migration after the year of perturbation (sample patches represented by Figure 4c can only be used for the tree classification task in the years of 2021–2023 and the recovery time of the sample patches is 2021).

All the above processes of calculation and judgement were carried out using the GEE platform. Eventually, we obtained raster data ranging from 1986–2023, denoting the time of forest restoration in the study area (the pixel values were uniformly set to 1986–present for pixels where no disturbance occurrence was detected). Subsequently, in order to verify the accuracy of the algorithm in determining the occurrence of disturbance and recovery of forest pixels, we imported the data of 5368 sample patches into the 'Google Earth' software and recorded the actual recovery time of each sample patch by visually judging the actual growth condition of each sample patch through the roll-up window. Finally, this study used R^2 and RMSE to assess the degree of fit between the actual recovery time and the predicted recovery time.
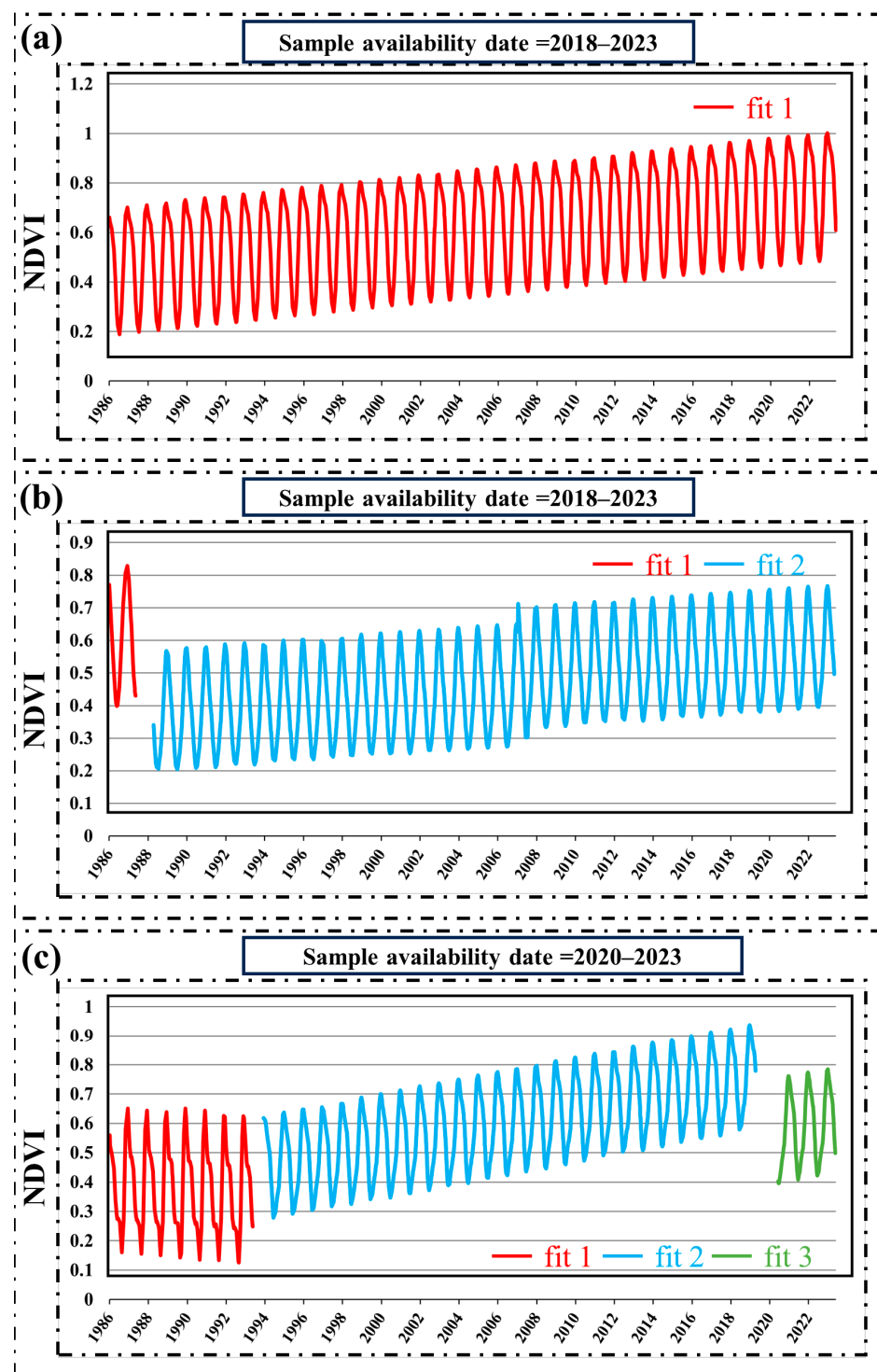
Figure 4. Examples of NDVI time series curves for sample patches under different disturbance scenarios. (**a**) represents an example of an undisturbed time-series curve within the entire time-series curve; (**b**) represents an example of a time-series curve that was disturbed prior to 2018; (**c**) represents an example of a time-series curve that was disturbed during the period of 2018–2023. Fit 1–3 represent each segment of the ongoing time-series curve. In subfigure (**a**), fit 1 represents the undisturbed NDVI time-series curve; in subfigure (**b**), fit 1 and fit 2 represent the NDVI time-series curves before and after being disturbed; and in subfigure (**c**), fits 1–3 represent the NDVI time-series curves before and after being subjected to multiple disturbances. Examples of NBR time-series curves for sample patches under different disturbance scenarios are shown in Figure S1 in the Supplementary Material.

3.2. Training of Classification Models

A complex and high-dimensional data processing problem is brought about by multi-temporal remote sensing images combined with vegetation index features and texture features, and machine learning algorithms are good coping tools [32,33]. The most common machine learning algorithms for forest monitoring tasks are the random forest algorithm (RF) and support vector machine algorithm (SVM) [9,34]. In recent years, the extreme gradient boosting (XGB) algorithm has gradually received much attention due to its excellent performance in forest monitoring tasks, obtaining higher classification accuracy than RF and SVM in crop classification and regional-scale forest species classification tasks [35–37]. So, in this study, RF, SVM, and XGB were used for the tree species classification task. Before proceeding with the classification model construction, in order to avoid the occurrence of the same sample being selected as both training and test samples during random sampling, this study firstly sets the 'random number' field for each sample patch, generates a random floating-point number directly from 0–1 in this field, and selects the samples with the value of this field greater than 0.3 for training samples and vice versa for test samples. In addition, in order to ensure the classification performance of the classification model when mapping tree species, we use 10% of the sample patches as modelling and regulating hyperparameter samples. After constructing the classification model based on the modeling sample, this study uses the grid search method for efficient optimal parameter selection. The hyperparameters of each algorithm are (1) the two most important hyperparameters constraining the classification performance of the RF algorithm: the number of decision trees (n_tree) and the maximum depth of each tree (max_depth) and (2) the penalty coefficients (C) and the kernel functions that are more helpful for constructing a robust and accurate SVM classification model. (3) The number of gradient trees (nrounds), the learning rate (eta), and the maximum tree depth (max_depth) are exceptionally important for improving the classification performance of the XGB algorithm. The optimal hyperparameters for each classification model are shown in Table 3. In this study, the RF and SVM algorithms are implemented in GEE and the XGB algorithm is implemented in Python 3.7.0.

Table 3. Optimal hyperparameters corresponding to each classification algorithm.

Classification Algorithms	Hyperparameters	Parameter Range	Optimal Hyperparameters
RF	n_tree	0–500	220
	max_depth	0–50	11
SVM	C	0.1–100	1
	Kernel	RBF, Linear, Poly, Sigmoid	RBF
XGB	nrounds	0–500	150
	Eta	0.001–1	0.036
	max_depth	0–50	7

3.3. Accuracy Evaluation

In this study, the following metrics were used to evaluate the accuracy of the classification model in the classification task: (1) confusion matrix, (2) overall accuracy, (3) Kappa coefficient, (4) user accuracy, and (5) producer accuracy.

4. Results

4.1. Results of Interannual Migration of the Sample

After analyzing and processing the results obtained by the CCDC change detection algorithm, we obtained the results of the detection of forest recovery time in the study area (Figure 5a). We can see that most of the forests in the mountainous areas in the southwest and northeast, where accessibility is poor and the intensity of development in the area is weak, have not been disturbed, and the forests as a whole have a long growth time. This suggests that the sample patches within this region can be used for inter-annual sample migration from 2018–2023. Conversely, in the central and western regions with higher

levels of urban development, forest disturbance occurs more severely, and some forests are disturbed in 2018–2023, which also represents that some samples within this region can only be migrated for a limited number of years. Through the visual interpretation of historical images (the specific judgement method is shown in Figure 5b–d), we analyzed the prediction accuracy of the change detection algorithm in determining the forest recovery time. The final result showed that the predicted forest restoration time has an accuracy of 0.8303, with an RMSE of 4.64 (Figure 6a). This favorably confirms the remarkable performance of the CCDC algorithm in determining forest disturbance and recovery. By analyzing the image elements corresponding to all sample patches, we obtained the number of available samples for each year from 2018–2023 (Figure 6b). The number of samples corresponding to each species in each year is detailed in Tables S1–S5. As can be seen from the tables, in almost all the years, the samples that could not be subjected to inter-annual migration were mainly from Eucalyptus and Pinus sylvestris. This is due to the fact that these two tree species are the main source of supply for forest management segments such as wood processing and paper making in the study area. Frequent felling rotations resulted in significant changes in their spectral differences during 2018–2022, which could easily be determined by the CCDC algorithm as forest disturbances occurring. The number of samples in each year, on the other hand, remained relatively stable, with small differences in the number of samples between tree species and a more balanced ratio. The influence of sample imbalance on the subsequent classification results was eliminated.
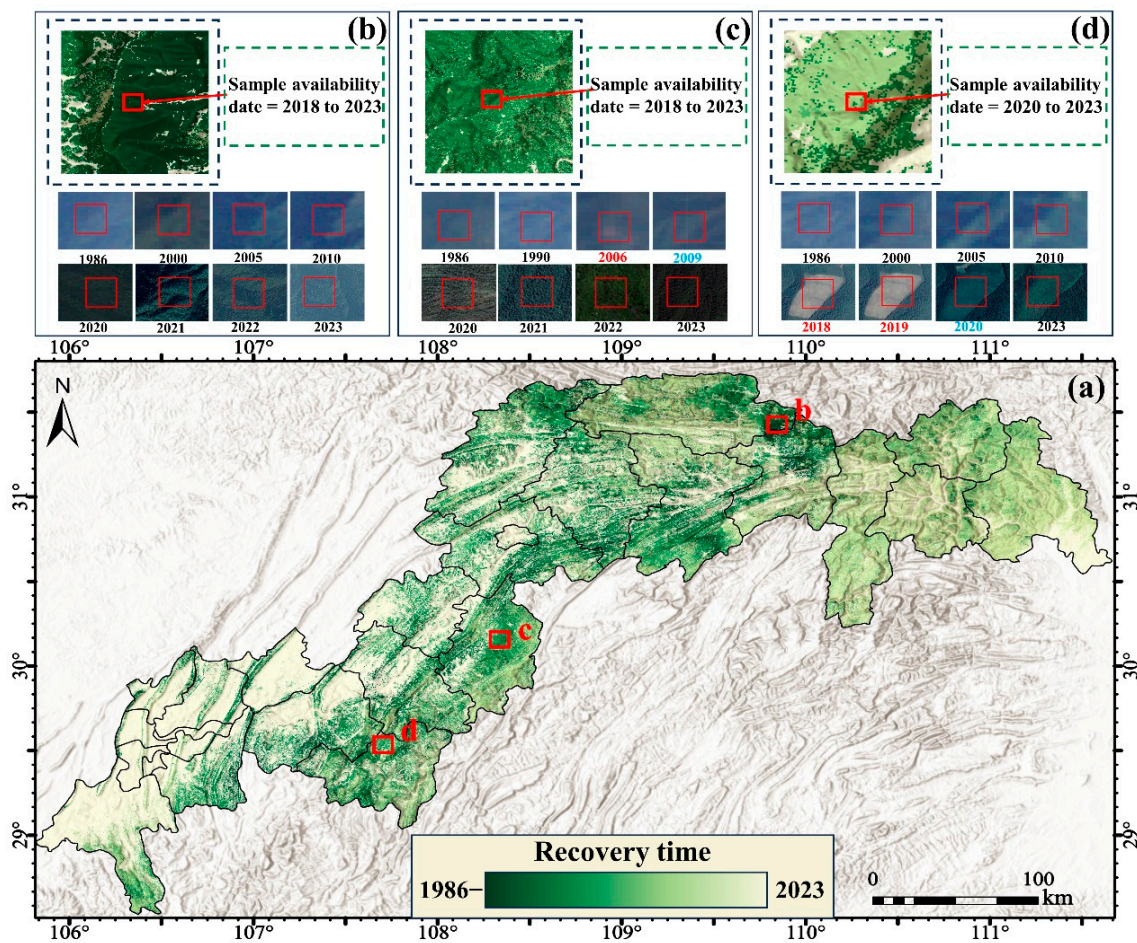

Figure 5. Calculations of forest recovery time in the study area and some examples thereof. (**a**) represents the distribution of forest recovery time calculation results. (**b**–**d**) represent the forest restoration time under different disturbances. The red font in the figure represents the occurrence of disturbance, and the blue font represents the completion of forest restoration.
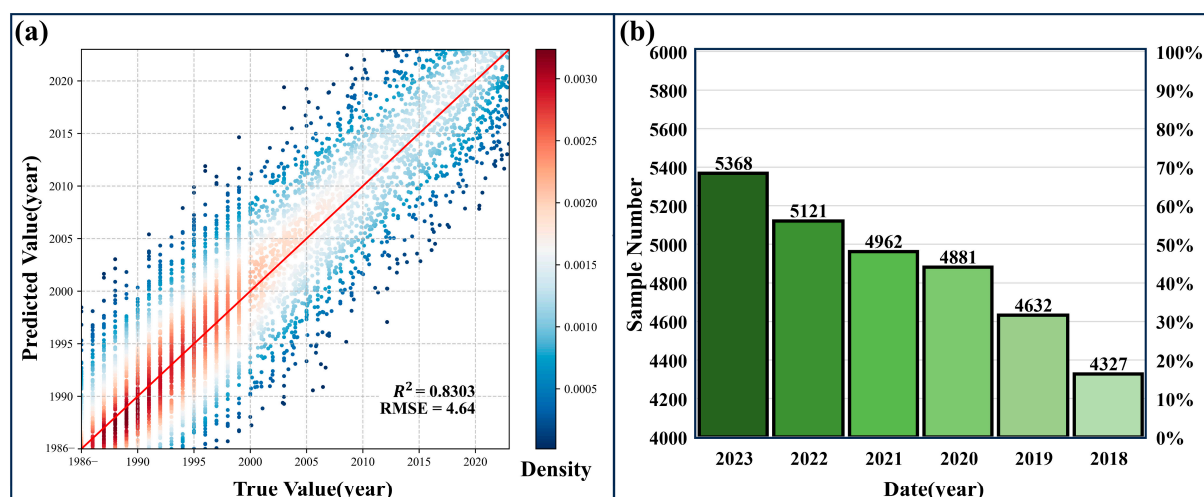
Figure 6. Accuracy validation of forest restoration time. (**a**) represents the degree of fit between actual and predicted restoration times; (**b**) represents the number of samples available during 2018–2023.

4.2. Comparison of the Accuracy of Classification Algorithms

As can be seen from Figure 7, all classification models had an overall accuracy higher than 70% in 2018–2023, with Kappa coefficients exceeding 0.6. Classification models based on the XGB algorithm achieved the highest classification accuracy in 2018–2023, with classification accuracies exceeding 80% in almost all the years and Kappa coefficients exceeding 0.75. The RF algorithm also exhibits strong classification performance, with the accuracy of all classification models exceeding 75% and Kappa coefficients greater than 0.7. The SVM algorithm, on the other hand, exhibits the poorest classification accuracy, with the accuracy of each classification model ranging from 71–75% and the Kappa coefficients ranging from 0.63 to 0.66. In addition, due to the difference between the number of available samples in each year, the classification accuracy varied from year to year. The year 2023 had the highest number of available samples and the highest classification accuracy for each model. It is worth mentioning that the XGB algorithm showed the strongest classification performance in the year 2023, achieving a classification accuracy of 88.05% with a Kappa coefficient of 0.8492. Since the XGB algorithm performed much better than the other algorithms in this study, the subsequent analysis will be centered on the classification results obtained by the XGB algorithm.

4.3. Dominant Tree Species Map Based on XGB Algorithm

Using each classification model based on the XGB algorithm to predict the forest species in the study area from 2018 to 2023, we obtained the distribution maps of forest dominant species in the study area (Figure 8). As we can see in the figure, Eucalyptus has the widest distribution in the study area, accounting for more than 40% of the total forest area, with a maximum of 44% in 2018, and generally showing a slight decreasing trend during the study period. Spatially, eucalypts are mainly distributed in the central and southwestern parts of the study area. It is followed by horsetail pine, mainly in the north and north-west. Fir is mainly found in the central region. Finally, cypress and wetland pine are mainly found in the southern and northeastern parts of the study area. The distribution of forest-dominant species in the study area was relatively stable from 2018 to 2023, benefitting from the government's strict forest protection policy.
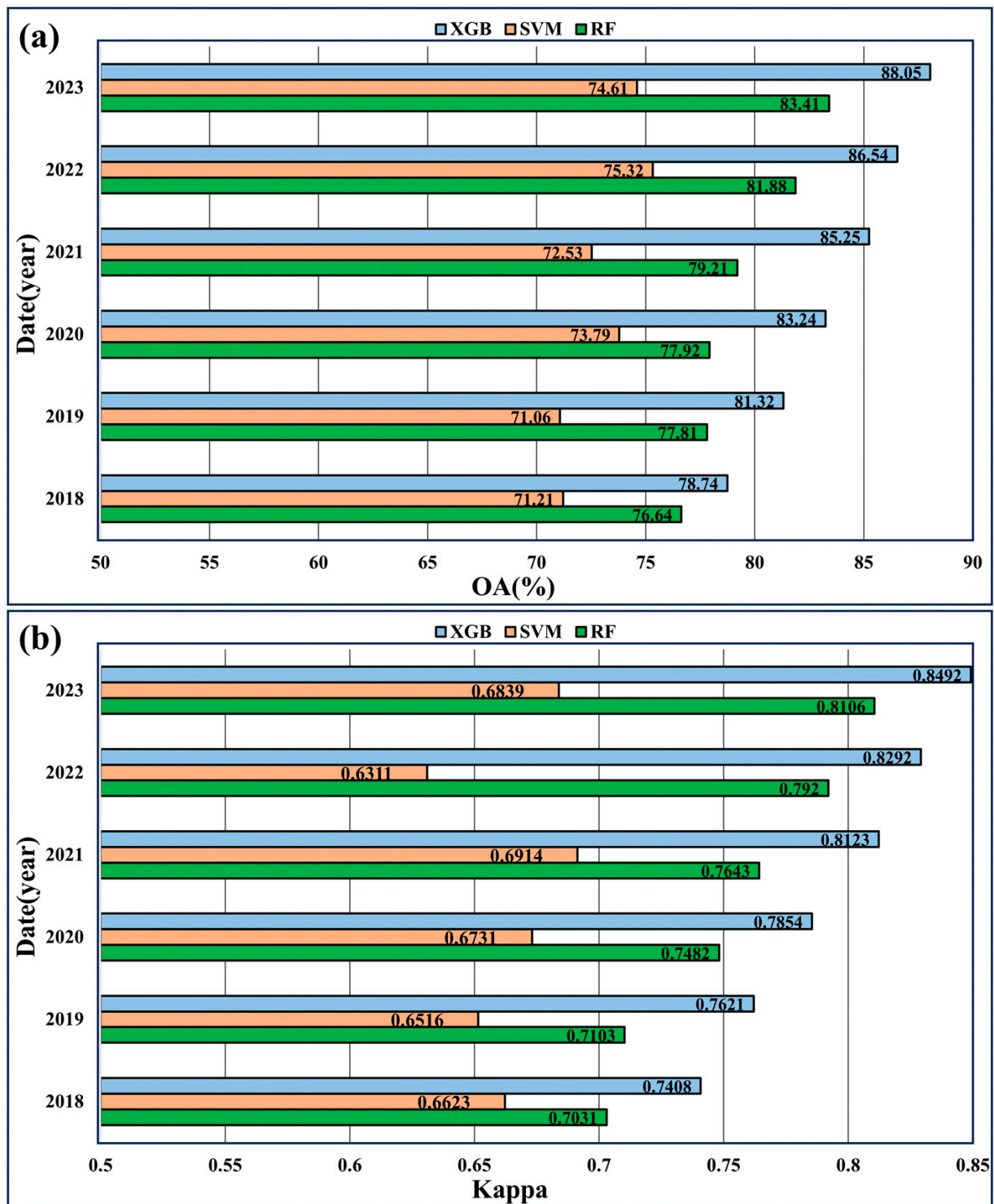
(a)

(b)

Figure 7. Classification accuracy of each classification model during 2018–2023. (**a**) represents the overall accuracy, and (**b**) represents the Kappa coefficient.
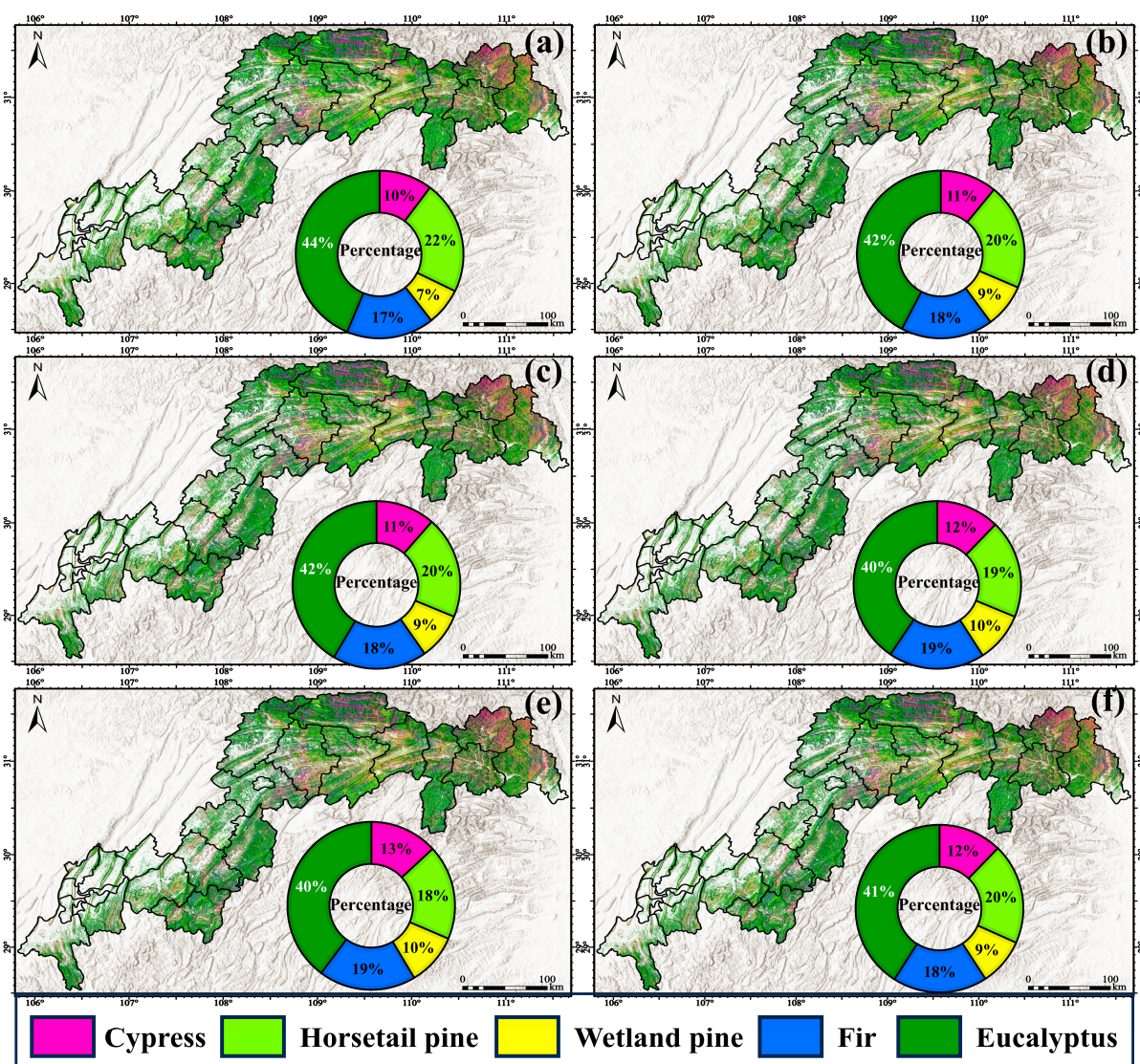
Figure 8. Distribution of dominant tree species based on the XGB algorithm 2018–2023. (**a**–**f**) represent the corresponding forest dominance tree maps for 2018–2023, respectively.

In the classification results, the classification accuracy of each tree species appeared to be quite different (e.g., Figure 9), with eucalyptus obtaining the highest PA and UA, which exceeded 85% in all years (it has the producer accuracy of 88–94% and the user accuracy of 85–94%). On the other hand, wetland pine showed more classification errors, which was the main source of error limiting the overall accuracy. In 2018–2023, wetland pine had a producer accuracy of 66–88% and a user accuracy of 69–84%. Its misclassification is mainly related to cypress and fir. This may be due to the mixed planting of these species in the study area. The rest of the tree species showed better classification accuracy (the producer accuracy of 72–88% and the user accuracy of 78–91%). As can be seen from Figure 10, the samples at the diagonal have an absolute numerical advantage. This represents that misclassification phenomena occur less frequently in the accuracy validation of our classification model. On the whole, our classification model shows excellent performance in the tree species classification task, and the finally obtained forest dominant species map has strong scientific significance.
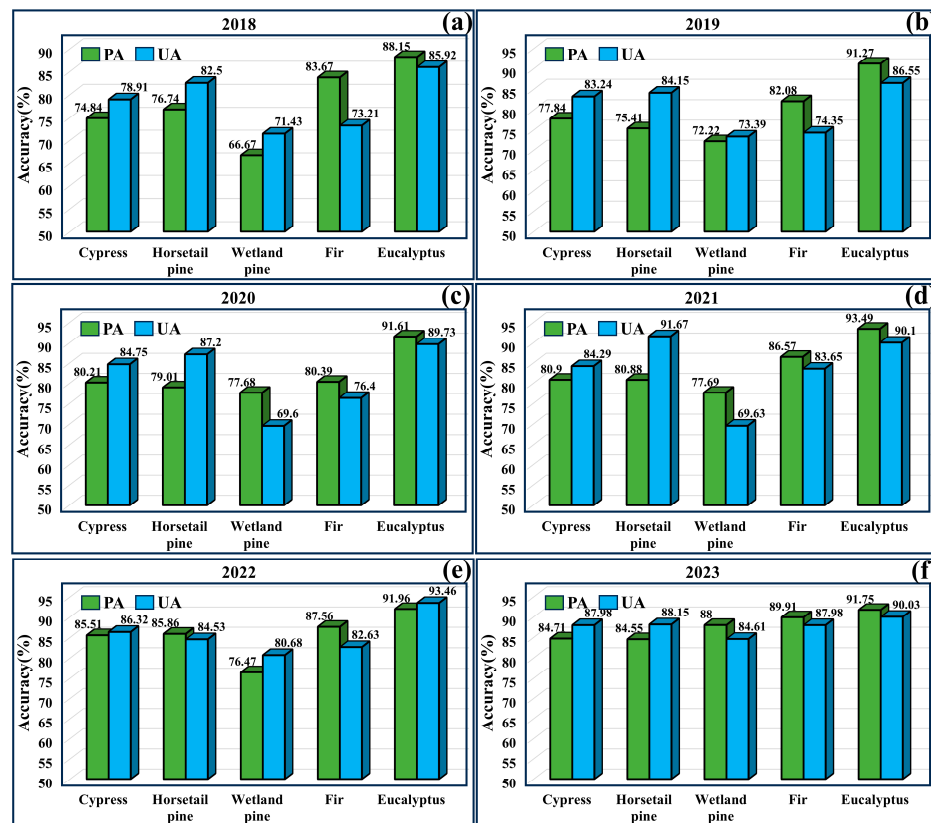
Figure 9. Producer accuracy and user accuracy for each tree species based on the XGB algorithm for 2018–2023. PA stands for producer accuracy, and UA stands for user accuracy. (**a**–**f**) represent the corresponding producer accuracy and user accuracy for 2018–2023, respectively.
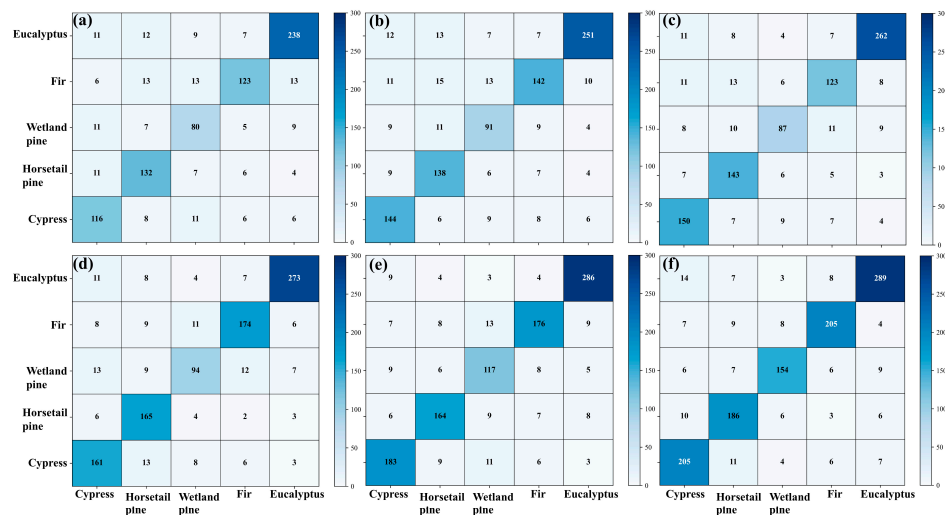

Figure 10. Confusion matrix based on the XGB algorithm. (**a**–**f**) represent the corresponding confusion matrices for 2018–2023, respectively.

4.4. Feature Importance Assessment Based on XGB Algorithm

In order to understand the contribution of each classification feature in the classification task, the mean decrease in the Gini (MDG) metric is used to assess the importance of classification features. As can be seen from Figure 11, the importance of each classification feature exhibits different distributions in both temporal and spectral perspectives. Temporally, most of the classification features showed persistent importance, such as B8, NDVI, NDWI, SAVI, and entropy, not just a single time step. Among them, the most notable one is

NDVI, which showed persistent and strong importance in all the images from 2018–2023. Spectrally, NDVI has the strongest contribution in the classification, followed by B8, B11, and B12 in the spectral bands. Among the textural features, entropy is of the highest importance. The above results proved that (1) features tend to show persistent importance in the classification model during the classification process. This also confirms the advantage of multi-temporal data over single-temporal data in the classification task (some of the features that show strong importance can be consistently useful in the classification task). (2) The excellent performance of vegetation indices in classification deserves to be noticed. However, low variable importance does not mean unimportant. Low variable importance can also be the result of a high correlation with highly important characteristics.
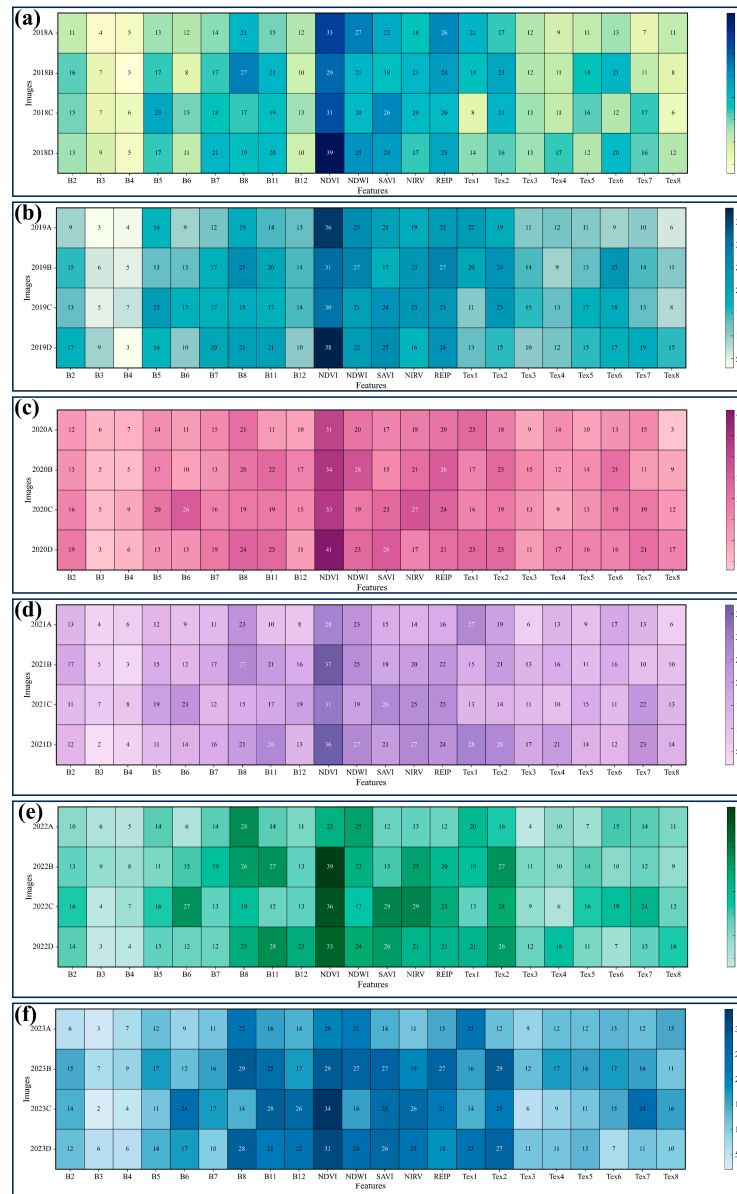
Figure 11. The feature importance of each classified feature in 2018–2023. (**a**–**f**) represent the feature importance of the categorical features in sequence during 2018–2023. A–D in the Y-axis labels represent the synthetic remote sensing images in each season. Tex1-8 in the X-axis labels represent the following textural features, respectively: energy, entropy, correlation, inverse difference moment, inertia, cluster shade, cluster prominence, and correlation.

5. Discussion

In recent years, the use of remote sensing for forest monitoring tasks has received increasing attention. However, attention has tended to focus on the classification of forest types in a single year. Multi-year mapping of forest species is essential for making forest management decisions. However, it is exceptionally difficult to carry out multi-year mapping at the species level due to data limitations. In this study, forest disturbance monitoring was carried out in each image element of the study area through the CCDC change detection algorithm. The samples in which no disturbance phenomenon occurred in the patches were migrated to 2018–2022, and based on the migrated samples, the dominant tree species map of forests in the Three Gorges Reservoir area from 2018–2023 was successfully produced, and a high classification accuracy was achieved. Among them, the XGB algorithm for the year 2023 showed the strongest classification performance and achieved 88.05% classification accuracy with a Kappa coefficient of 0.8492.

Accurate identification of invariant samples can be achieved using the CCDC change detection algorithm. In this study, the CCDC algorithm showed excellent identification results, where R^2 was 0.8303, with an RMSE of 4.64. However, the recovery state of the forests after disturbance is highly correlated with the actual function of the forests in the region. Take eucalyptus as an example: as a representative of fast-growing timber forests, it has a rotational felling period of 3–5 years. This means that eucalypts are cut down immediately after 3–5 years of growth and then quickly transplanted by hand [5]. On the other hand, other tree species may exist in a state of idleness for a period of time after being disturbed. During this period, the land cover type of the area may be bare soil or grassland, and the period between disturbance and restoration to the forest remains full of uncertainty. Although we set a series of rules to eliminate the misjudgment of the algorithm, this does not guarantee the timely implementation of capturing the time of forest restoration and may cause the algorithm to make errors in judging the time of forest restoration. This issue deserves attention in subsequent studies.

This study successfully mapped the dominant tree species for 2018–2023 using seasonal synthetic images for each year. However, due to rapid climatic changes, the limited seasonal synthetic imagery is not sufficient to capture the complete climatic changes, which may lead to the loss of information on tree species growth and thus result in classification errors. In subsequent studies, it is necessary to use more intensive remote sensing data for tree species classification. For example, Hemmerling et al. [38], in a study on the classification of temperate tree species, indicated that time series data could provide better climatic information of tree species in classification, thus enhancing the classification effect. Likewise, Huang et al. [39], in the identification of tree species in plantation forests using deep learning methods, indicated that the climatic information contained in time-series data is essential for improving classification accuracy. Similarly, NDVI consistently showed strong importance in the feature importance evaluation part of our study. This laterally confirms the possibility of more intensive remote sensing data in improving classification accuracy. However, the acquisition of time series data can be fraught with difficulties due to the study area of more than 60,000 km^2 in this study. Fortunately, there are some studies that address this problem by interpolating time-series data or data synthesis. Blickensdörfer et al. [40] achieved the synthesis of national scale time series data based on a weighted convolutional filter by performing a national scale classification of tree species considered. Hermosilla et al. [17] provided us with another idea: seamless annual data covering the whole of Canada was achieved based on the best available pixel (BAP) method. These methods provide valuable inspiration and reference for our future research work.

In this study, the XGB algorithm achieved the highest classification accuracy in almost all years. This is highly similar to the findings in previous related studies. This is because the gradient-boosting framework of the XGB algorithm in efficiently processing datasets with high-dimensional feature spaces ensures that each new model attempts to correct the errors present in the previous model. Its regularization parameter controls the complexity of the model and prevents overfitting. Thus, it can ensure that the classification results of

the classification model at unknown image elements are more reliable. However, there are fewer studies related to it, and its effectiveness deserves further exploration.

Tree species sample data were the basis of this study. The main sources of sample data for this study are (1) National Forest Inventory (NFI) data and (2) field measurements based on expert knowledge, respectively. The final classification results demonstrate high classification accuracy. However, the accuracy of the data is still questionable. The NFI data may be subject to some uncertainties due to their long aggregation date: the plot may be subject to forest disturbances, etc., during the period of time after the tree species information has been recorded and up to the time when the whole forest inventory task is completed. Furthermore, field measurement data are highly dependent on the professionalism of the forester. Subjective judgments of tree species classifications may be a potential contributor to classification errors.

6. Conclusions

In this study, to address the problem of missing annual ground sample data that may exist in the remote sensing monitoring of forest-dominant tree species, we used the CCDC algorithm to carry out inter-annual migration of sample data and achieved a high migration accuracy. On this basis, we comprehensively compared the accuracies of various machine learning algorithms in the classification of forest-dominant tree species in the Three Gorges Reservoir area of China and adopted the XGB algorithm, which has the highest classification accuracy, to carry out the mapping of forest dominant tree species in the study area from 2018–2023, thus grasping the area change and spatial distribution of forest dominant tree species in the Three Gorges Reservoir area of China over the past 6 years. The main findings are as follows:

1. The CCDC algorithm shows excellent performance in sample migration. The final results obtained have high accuracy, with R^2 of 0.8303 and RMSE of 4.64. The XGB algorithm has an absolute advantage.
2. The absolute advantage of the XGB algorithm. The classification model based on the XGB algorithm shows significant classification advantages every year, with classification accuracies above 80% and Kappa coefficients higher than 0.75 in almost all years. In particular, the XGB algorithm in 2023 shows the strongest classification performance, achieving a classification accuracy of 88.05% and a Kappa coefficient of 0.8492.
3. Continued importance of classification features. In this study, it was found that most of the features showed sustained importance in the feature importance assessment based on MDG metrics, such as B8, NDVI, NDWI, SAVI, and entropy. Among them, the most noteworthy one is NDVI, which showed sustained and strong importance from 2018 to 2023.

Our research proposed a work path that combines sample migration and machine learning for forest dominant species classification, which is a useful exploration of temporal domain migration learning.

References

1. Gamfeldt, L.; Snäll, T.; Bagchi, R.; Jonsson, M.; Gustafsson, L.; Kjellander, P.; Ruiz-Jaen, M.C.; Fröberg, M.; Stendahl, J.; Philipson, C.D. Higher levels of multiple ecosystem services are found in forests with more tree species. *Nat. Commun.* **2013**, *4*, 1340. [CrossRef] [PubMed]
2. Vihervaara, P.; Auvinen, A.-P.; Mononen, L.; Törmä, M.; Ahlroth, P.; Anttila, S.; Böttcher, K.; Forsius, M.; Heino, J.; Heliölä, J. How essential biodiversity variables and remote sensing can help national biodiversity monitoring. *Glob. Ecol. Conserv.* **2017**, *10*, 43–59. [CrossRef]
3. Lehtomäki, J.; Tuominen, S.; Toivonen, T.; Leinonen, A. What data to use for forest conservation planning? A comparison of coarse open and detailed proprietary forest inventory data in Finland. *PLoS ONE* **2015**, *10*, e0135926.
4. Franklin, S.E. *Remote Sensing for Sustainable Forest Management*; CRC Press: Boca Raton, FL, USA, 2001.
5. Li, Y.; Liu, X.; Liu, M.; Wu, L.; Zhu, L.; Huang, Z.; Xue, X.; Tian, L. Historical Dynamic Mapping of Eucalyptus Plantations in Guangxi during 1990–2019 Based on Sliding-Time-Window Change Detection Using Dense Landsat Time-Series Data. *Remote Sens.* **2024**, *16*, 744. [CrossRef]
6. McRoberts, R.E.; Tomppo, E.O.; Næsset, E. Advances and emerging issues in national forest inventories. *Scand. J. For. Res.* **2010**, *25*, 368–381. [CrossRef]
7. Barrett, F.; McRoberts, R.E.; Tomppo, E.; Cienciala, E.; Waser, L.T. A questionnaire-based review of the operational use of remotely sensed data by national forest inventories. *Remote Sens. Environ.* **2016**, *174*, 279–289. [CrossRef]
8. Wulder, M.A.; Kurz, W.A.; Gillis, M. National level forest monitoring and modeling in Canada. *Prog. Plan.* **2004**, *61*, 365–381. [CrossRef]
9. Fassnacht, F.E.; Latifi, H.; Stereńczak, K.; Modzelewska, A.; Lefsky, M.; Waser, L.T.; Straub, C.; Ghosh, A. Review of studies on tree species classification from remotely sensed data. *Remote Sens. Environ.* **2016**, *186*, 64–87. [CrossRef]
10. Axelsson, A.; Lindberg, E.; Reese, H.; Olsson, H. Tree species classification using Sentinel-2 imagery and Bayesian inference. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *100*, 102318. [CrossRef]
11. Waser, L.T.; Küchler, M.; Jütte, K.; Stampfer, T. Evaluating the potential of WorldView-2 data to classify tree species and different levels of ash mortality. *Remote Sens.* **2014**, *6*, 4515–4545. [CrossRef]
12. Liu, L.; Coops, N.C.; Aven, N.W.; Pang, Y. Mapping urban tree species using integrated airborne hyperspectral and LiDAR remote sensing data. *Remote Sens. Environ.* **2017**, *200*, 170–182. [CrossRef]
13. Immitzer, M.; Atzberger, C.; Koukal, T. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data. *Remote Sens.* **2012**, *4*, 2661–2693. [CrossRef]
14. Cheng, K.; Chen, Y.; Xiang, T.; Yang, H.; Liu, W.; Ren, Y.; Guan, H.; Hu, T.; Ma, Q.; Guo, Q. A 2020 forest age map for China with 30 m resolution. *Earth Syst. Sci. Data* **2024**, *16*, 803–819. [CrossRef]
15. Koskinen, J.; Leinonen, U.; Vollrath, A.; Ortmann, A.; Lindquist, E.; d'Annunzio, R.; Pekkarinen, A.; Käyhkö, N. Participatory mapping of forest plantations with Open Foris and Google Earth Engine. *ISPRS J. Photogramm. Remote Sens.* **2019**, *148*, 63–74. [CrossRef]
16. Cook-Patton, S.C.; Leavitt, S.M.; Gibbs, D.; Harris, N.L.; Lister, K.; Anderson-Teixeira, K.J.; Briggs, R.D.; Chazdon, R.L.; Crowther, T.W.; Ellis, P.W.; et al. Mapping carbon accumulation potential from global natural forest regrowth. *Nature* **2020**, *585*, 545–550. [CrossRef]
17. Hermosilla, T.; Bastyr, A.; Coops, N.C.; White, J.C.; Wulder, M.A. Mapping the presence and distribution of tree species in Canada's forested ecosystems. *Remote Sens. Environ.* **2022**, *282*, 113276. [CrossRef]
18. Wulder, M.A.; Loveland, T.R.; Roy, D.P.; Crawford, C.J.; Masek, J.G.; Woodcock, C.E.; Allen, R.G.; Anderson, M.C.; Belward, A.S.; Cohen, W.B. Current status of Landsat program, science, and applications. *Remote Sens. Environ.* **2019**, *225*, 127–147. [CrossRef]
19. White, J.C.; Gómez, C.; Wulder, M.A.; Coops, N.C. Characterizing temperate forest structural and spectral diversity with Hyperion EO-1 data. *Remote Sens. Environ.* **2010**, *114*, 1576–1589. [CrossRef]

20. Yin, H.; Khamzina, A.; Pflugmacher, D.; Martius, C. Forest cover mapping in post-Soviet Central Asia using multi-resolution remote sensing imagery. *Sci. Rep.* **2017**, *7*, 1375. [CrossRef]

21. Yang, B.; Wu, L.; Liu, M.; Liu, X.; Zhao, Y.; Zhang, T. Mapping Forest Tree Species Using Sentinel-2 Time Series by Taking into Account Tree Age. *Forests* **2024**, *15*, 474. [CrossRef]

22. Grabska, E.; Hostert, P.; Pflugmacher, D.; Ostapowicz, K. Forest stand species mapping using the Sentinel-2 time series. *Remote Sens.* **2019**, *11*, 1197. [CrossRef]

23. Farwell, L.S.; Gudex-Cross, D.; Anise, I.E.; Bosch, M.J.; Olah, A.M.; Radeloff, V.C.; Razenkova, E.; Rogova, N.; Silveira, E.M.; Smith, M.M. Satellite image texture captures vegetation heterogeneity and explains patterns of bird richness. *Remote Sens. Environ.* **2021**, *253*, 112175. [CrossRef]

24. Adams, B.; Iverson, L.; Matthews, S.; Peters, M.; Prasad, A.; Hix, D.M. Mapping forest composition with landsat time series: An evaluation of seasonal composites and harmonic regression. *Remote Sens.* **2020**, *12*, 610. [CrossRef]

25. Ahlswede, S.; Schulz, C.; Gava, C.; Helber, P.; Bischke, B.; Förster, M.; Arias, F.; Hees, J.; Demir, B.; Kleinschmit, B. TreeSatAI Benchmark Archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth Syst. Sci. Data* **2023**, *15*, 681–695. [CrossRef]

26. Zhang, H.K.; Roy, D.P. Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification. *Remote Sens. Environ.* **2017**, *197*, 15–34. [CrossRef]

27. Fekri, E.; Latifi, H.; Amani, M.; Zobeidinezhad, A. A Training Sample Migration Method for Wetland Mapping and Monitoring Using Sentinel Data in Google Earth Engine. *Remote Sens.* **2021**, *13*, 4169. [CrossRef]

28. Yang, J.; Huang, X. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth Syst. Sci. Data* **2021**, *13*, 3907–3925. [CrossRef]

29. Zhu, Z.; Woodcock, C.E. Continuous change detection and classification of land cover using all available Landsat data. *Remote Sens. Environ.* **2014**, *144*, 152–171. [CrossRef]

30. Xiao, Y.; Wang, Q.; Tong, X.; Atkinson, P.M. Thirty-meter map of young forest age in China. *Earth Syst. Sci. Data* **2023**, *15*, 3365–3386. [CrossRef]

31. Du, Z.; Yu, L.; Yang, J.; Xu, Y.; Chen, B.; Peng, S.; Zhang, T.; Fu, H.; Harris, N.; Gong, P. A global map of planting years of plantations. *Sci. Data* **2022**, *9*, 141. [CrossRef]

32. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geosciences and remote sensing. *Geosci. Front.* **2016**, *7*, 3–10. [CrossRef]

33. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817. [CrossRef]

34. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]

35. Grabska, E.; Frantz, D.; Ostapowicz, K. Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the Polish Carpathians. *Remote Sens. Environ.* **2020**, *251*, 112103. [CrossRef]

36. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Kalogirou, S.; Wolff, E. Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application. *GISci. Remote Sens.* **2018**, *55*, 221–242. [CrossRef]

37. Zhong, L.; Hu, L.; Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* **2019**, *221*, 430–443. [CrossRef]

38. Hemmerling, J.; Pflugmacher, D.; Hostert, P. Mapping temperate forest tree species using dense Sentinel-2 time series. *Remote Sens. Environ.* **2021**, *267*, 112743. [CrossRef]

39. Huang, Z.; Zhong, L.; Zhao, F.; Wu, J.; Tang, H.; Lv, Z.; Xu, B.; Zhou, L.; Sun, R.; Meng, R. A spectral-temporal constrained deep learning method for tree species mapping of plantation forests using time series Sentinel-2 imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 397–420. [CrossRef]

40. Blickensdörfer, L.; Oehmichen, K.; Pflugmacher, D.; Kleinschmit, B.; Hostert, P. National tree species mapping using Sentinel-1/2 time series and German National Forest Inventory data. *Remote Sens. Environ.* **2024**, *304*, 114069. [CrossRef]