Dominant Tree Species Extraction Algorithm Using Machine Learning Based on Multi-temporal High-resolution Data

Guo H.Y. [1,2], Liang W.[3], Zhang Z.D.[4], Wang S.H.[3], Xu M.[3] and Cao C.X. [3,*]

[1]Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, China

[2] University of Chinese Academy of Sciences, China

[3] Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, China

[4] Academy of Forestry Inventory and Planning,National Forestry and Grassland Administration,China

* caocx@aircas.ac.cn

***Abstract*:** *Timely and accurate mapping of forest types is essential for forest resource inventories, providing critical support for forest management, conservation biology, and ecological restoration. Tree species classification plays a significant role in promoting sustainable forest management and protecting ecological environments. In this study, Sentinel-1 and Sentinel-2 data were utilized to classify six dominant tree species in the Chengde and Beijing regions: Larix spp., Pinus tabulaeformis, Platycladus spp., Quercus L., Betula spp., and Betula platyphylla. To effectively capture temporal variations, data were acquired in March, June, September, and December 2020, and a variety of features were extracted, including Sentinel-1 bands, spectral indices, Sentinel-2 bands, spectral indices, texture features, and topographic variables. Forest inventory data were employed as sample information to explore the optimal combination of input variables. In total, 1,519 field survey samples were used to construct training and testing datasets. The classification process employed both the Random Forest (RF) and XGBoost algorithms, with model performance evaluated using the out-of-bag (OOB) score and cross-validation methods. Results indicated that the highest classification accuracy for the RF model (78.07%, kappa = 0.691) was achieved when Sentinel-1, Sentinel-2 indices, Sentinel-2 texture features, and digital elevation model (DEM) data were used as input variables. For the XGBoost model, the highest classification accuracy (81.25%, kappa = 0.737) was obtained when Sentinel-1, Sentinel-2 bands, Sentinel-2 indices, Sentinel-2 texture features, and DEM data were incorporated. In the study area, Quercus spp. was the dominant tree species, covering 66% of the area, followed by Pinus tabulaeformis, which occupied 19.7%. The results demonstrate the potential of using Sentinel-1 and Sentinel-2 data for tree species classification, and highlight the effectiveness of machine learning algorithms in this application. This study underscores the capability of combined synthetic aperture radar (SAR) and optical data for large-scale tree species classification and suggests significant implications for forest monitoring and management..*

Keywords:tree species, sentinel-1/2, machine learning, large scale mapping

1. Introduction

Forests, as a key component of terrestrial ecosystems, play an essential role in maintaining biodiversity, regulating the global climate, sustaining ecological balance, and contributing to the global carbon cycle (Sabins Jr & Ellis, 2020). Effective and scientific management of forest resources requires a comprehensive understanding of forest types, quantities, and

spatial distributions. Accurate identification of tree species is fundamental to the sustainable utilization and conservation of forest resources (Fassnacht et al., 2016). However, traditional methods of forest resource inventory, which largely rely on ground-based surveys, are often constrained by high costs, time consumption, labor intensity, and limited spatial coverage. These shortcomings hinder their ability to meet the needs of contemporary forest resource management, particularly in providing detailed spatial information on forest stand types.In contrast, remote sensing technology offers large-scale spatial data and, compared to traditional ground-based forest assessments, provides significant advantages, including macroscopic perspective, rapid data acquisition, and cost-efficiency(Sabins Jr & Ellis, 2020). The integration of satellite remote sensing in forest resource surveys enhances the precision and efficiency of forest resource assessment and management. Tree species identification using satellite imagery is a critical tool for quantitatively estimating forest leaf area index, carbon storage, and biomass. Additionally, it supports efforts to address key "carbon cycle" challenges, such as the determination of "carbon sources" and "carbon sinks" (Dalponte et al., 2012). This technological advancement is instrumental in advancing forest resource monitoring and promoting sustainable forest management practices.

In recent years, optical, LiDAR, and radar remote sensing data have been successfully employed in forest tree species classification, yielding promising outcomes (Fassnacht et al., 2016). Remote sensing imagery with medium spatial resolution, characterized by high temporal and spectral resolution, extensive spatial coverage, a short revisit cycle, and substantial data accumulation over extended periods, represents a valuable data source for enhancing the accuracy of forest type extraction. It is particularly well-suited for large-scale regional applications (Michałowska & Rapiński, 2021). However, meteorological factors can introduce considerable variability in tree species classification when relying on single-date remote sensing data. Therefore, the use of multi-temporal remote sensing data has proven to significantly improve the accuracy of dominant tree species identification by leveraging the temporal variability of multispectral data(Persson et al., 2018). With the rapid advancement of hyperspectral remote sensing, hyperspectral imagery, which offers enriched spectral information, has been demonstrated as a highly effective tool for tree species classification in numerous studies (Dalponte et al., 2012). However, challenges associated with the acquisition of hyperspectral data, coupled with the complexity of data processing, currently limit the application of hyperspectral remote sensing for tree species classification to smaller forest areas. Simultaneously, advancements in aerial platforms and unmanned aerial vehicles (UAVs) have spurred rapid developments in airborne LiDAR technology. Owing to its ability

to accurately capture the three-dimensional structural characteristics of trees, airborne LiDAR data has been widely utilized in the mapping of forest species composition.

In conclusion, various remote sensing data deployed on different platforms exhibit distinct strengths and limitations in tree species classification. Consequently, recent research has increasingly focused on the fusion of multi-source remote sensing data to enhance classification accuracy. This approach leverages the integration of data from different spatial and spectral resolutions, as well as multiple sensor platforms, to mitigate the limitations of single-source data. Compared to the challenges associated with acquiring and processing hyperspectral data for large-scale tree species distribution mapping, freely available Sentinel-1 and Sentinel-2 data are increasingly favored by researchers. Since its launch in 2015, Sentinel-2, with its red-edge bands and high spatial resolution, has proven to be highly effective for large-area vegetation monitoring. For example, Magnus Persson utilized Sentinel-2A to study tree species classification in Swedish forests, demonstrating its efficacy for this purpose (Michałowska & Rapiński, 2021). Radar data, characterized by its all-weather and day-and-night monitoring capabilities, is particularly valuable for classifying tropical and subtropical wetland vegetation(Michałowska & Rapiński, 2021). Onojeghuo et al. employed C-band Sentinel-1 data on the Google Earth Engine (GEE) platform for tree species classification in Canadian wetland reserves. By combining radar data with machine learning algorithms, they achieved high classification accuracy(Onojeghuo et al., 2021). A review of the existing literature indicates that the fusion of multispectral, radar, and topographic data yields superior classification results(Abdollahnejad & Panagiotidis, 2020). Therefore, this study utilizes Sentinel optical imagery, Sentinel radar data, and multi-temporal satellite imagery as primary data sources. Topographic data is incorporated as auxiliary information to investigate the optimal classification model using various machine learning algorithms under different data input scenarios. This approach aims to enhance the precision of tree species classification and contribute to advancements in forest resource monitoring.

In the domain of tree species identification and classification, object-based classification is a widely adopted approach for hyperspectral data, as it allows for the extraction of detailed tree species information(Franklin & Ahmed, 2018). With the continuous advancement of deep learning models, these methods have facilitated automatic feature extraction and end-to-end tree species classification by processing hyperspectral data(Fujimoto et al., 2019). However, such approaches are primarily designed for hyperspectral data, limiting their scalability and applicability for tree species classification across large geographical

areas(Fassnacht et al., 2016). In contrast, tree species classification based on the fusion of multi-source remote sensing data typically relies on machine learning techniques(Franklin & Ahmed, 2018). Commonly used models in this context include Random Forest (RF), and eXtreme Gradient Boosting (XGBoost).

In this study, multi-source remote sensing data, including band information, remote sensing indices, texture features, and topographic information, are processed and utilized as input for various classification models. These datasets are grouped and fed into different models to evaluate the effectiveness of each type of remote sensing data in tree species identification. By comparing the classification accuracies of these models, the study aims to determine the model that delivers the highest accuracy in tree species classification. Ultimately, this research seeks to produce a comprehensive tree species distribution map for the study area, which will provide valuable data for subsequent estimation of aboveground forest biomass. Furthermore, the results will offer critical insights to forestry departments, aiding in forest resource management and conservation efforts.

2. Materials and Methods

2.1 Study Area:

This study focuses on two regions: Beijing and Chengde. Beijing, situated in the North China Plain, spans from 115°42′E to 117°42′E longitude and 39°24′N to 41°36′N latitude, covering a total area of approximately16,400km ^{2} . Forest resources in Beijing are predominantly distributed in the mountainous areas to the west and north, with the primary tree species including Pinus tabuliformis, Larix spp., Platycladus orientalis, and Betula spp.(Li et al., 2015). According to the Beijing Statistical Yearbook, as of 2020, the forested area reached 848,000 hectares, with a forest coverage rate of 44.4%.Chengde, adjacent to Beijing, is located between 115°54′E and 119°15′E longitude and 40°11′N and 42°40′N latitude, encompassing an area of 39,500 km^2. As of 2020, Chengde's forest resources cover 2.37 million hectares, with a forest coverage rate of 60%, accounting for 35.7% of the total forest area in Hebei Province. The dominant tree species in Chengde are Pinus tabuliformis, Platycladus orientalis, Betula spp., and Quercus spp.(Ming et al., 2021).

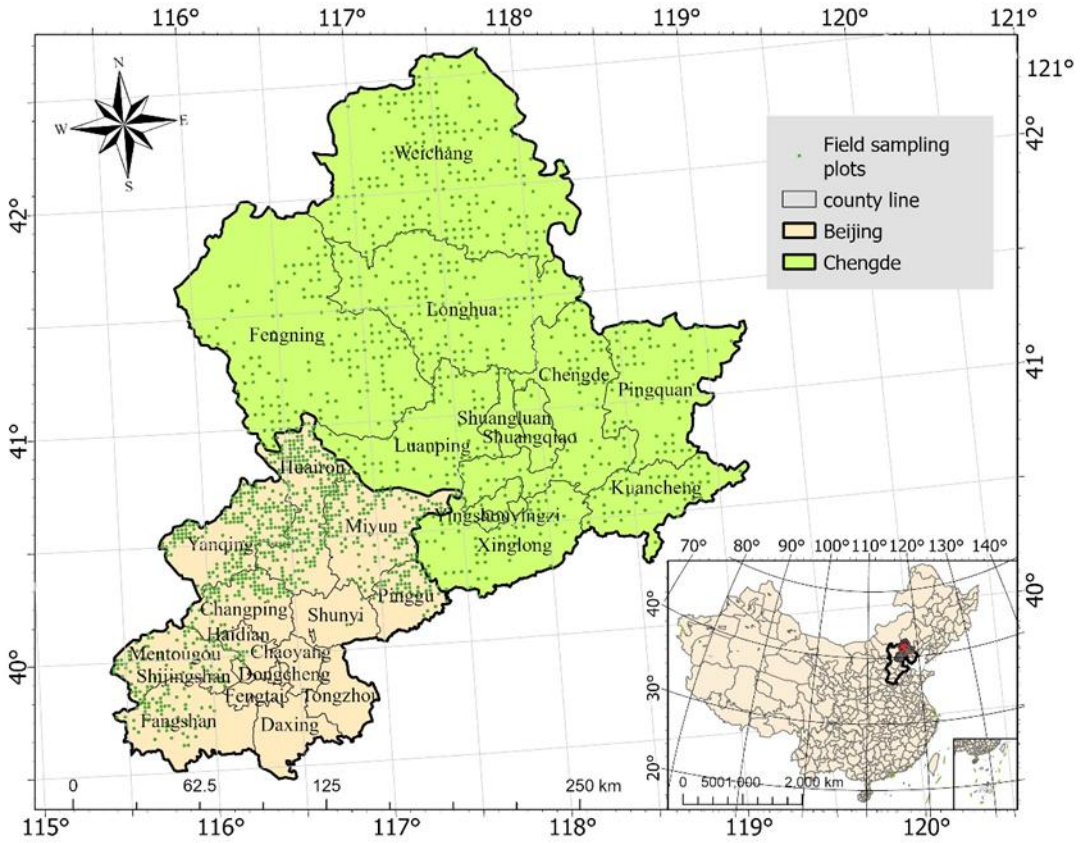

Figure 1: Sampling points in the study area and the sample plot.

2.2 Sentinel-1:

The Sentinel-1 satellites are equipped with C-band synthetic aperture radar (SAR) and offer four distinct imaging modes: Extra Wide Swath (EW), Strip Map Mode (SM), Wave Mode (WV), and Interferometric Wide Swath (IW). For the purposes of this study, Sentinel-1 IW mode Ground Range Detected (GRD) data are utilized, specifically focusing on the VV (Vertical-Vertical) and VH (Vertical-Horizontal) dual-polarization characteristics for subsequent analysis(Torres et al., 2012).

2.3 Sentinel-2:

The Sentinel-2 satellites, equipped with high-resolution multispectral imaging instruments, are specifically designed for terrestrial monitoring, providing detailed imagery of vegetation, soil, water bodies, as well as inland and coastal regions(Drusch et al., 2012). In this study, Sentinel-2 data were sourced from the Google Earth Engine (GEE) platform, encompassing the months of March, June, September, and December 2020, to represent seasonal variability. Initially, cloud masking techniques were applied to exclude cloudy pixels and ensure the retention of high-quality, cloud-free data(Coluzzi et al., 2018). Subsequently, monthly composite remote sensing images were generated using the median compositing method.

Vegetation indices were derived based on the available 12 spectral bands, with the specific indices utilized outlined in the corresponding table.

2.4 Field Data and Auxiliary Data:

Forest resource survey data provide an objective representation of ground conditions and the status of forest resources during the study period, offering crucial insights into the spatial distribution of vegetation and land cover types within the study area(Majasalmi et al., 2018). In this study, data from the ninth national forest resource inventory were utilized, which includes detailed records of dominant tree species within the sample plots. This inventory is conducted on a five-year cycle, with a total of 1,519 survey plots distributed across Beijing and Chengde, as shown in Figure 1. All survey data are based on 2020 records, aligning with the temporal scope of the satellite imagery employed in the classification process. Furthermore, Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) data were incorporated as auxiliary data, providing three topographic variables: elevation, slope, and aspect(Sesnie et al., 2008).

2.5 Data Processing:

Sentinel-1 data ('COPERNICUS/S1_GRD') in Interferometric Wide Swath (IW) mode were acquired from Google Earth Engine (GEE) for vertical-vertical (VV) and vertical-horizontal (VH) polarization backscatter coefficients for the months of March, June, September, and December 2020. From this dataset, three indices were computed: Backscatter Division, Backscatter Difference, and Backscatter Amplitude(Alexander et al., 2010).

For Sentinel-2 data ('COPERNICUS/S2_SR_HARMONIZED'), spectral bands B2, B3, B4, B5, B6, B7, and B8 were utilized. Eight vegetation indices were derived: Normalized Difference Vegetation Index (NDVI)(Madonsela et al., 2018), Modified Normalized Difference Water Index (MNDWI)(Sun et al., 2021), Normalized Difference Built-up Index (NDBI)(Dalponte et al., 2014), Enhanced Vegetation Index (EVI)(Arvor et al., 2011), Burn Severity Index (BSI)(Cornelis et al., 2010), Sentinel-2 Red-Edge Position Index (S2REP)(Bhattarai et al., 2022), Green Normalized Difference Vegetation Index (GNDVI)(Otsu et al., 2019), and Meris Terrestrial Chlorophyll Index (MTCI)(Choi et al., 2011). Additionally, texture features were extracted from the red-edge bands of Sentinel-2 using the Gray-Level Co-occurrence Matrix (GLCM)(Yang et al., 2019), resulting in eight texture metrics: Mean, Variance, Homogeneity, Contrast, Dissimilarity, Entropy, Angular Second Moment, and Correlation.In total, the data processing yielded 115 features for subsequent modeling: 20 features from Sentinel-1 (5 indices across 4 time points), 60 features from Sentinel-2 spectral bands (15 bands across 4 time points), 32 texture features from

Sentinel-2 (8 features across 4 time points), and 3 topographic features. These features will be employed in the development of models for classifying dominant tree species.

Number	Feature	Formula	Source
8	Polarization bands	VV,VH	Sentinel-1
4	Back scatter difference	VH-VV	Sentinel-1
4	Back scatter division	VH/VV	Sentinel-1
4	Back scatter amplitude	$\sqrt{VH^2 + VV^2}$	Sentinel-1
28	Spectral bands	Blue,Green,Red,Red-edge1, Red-edge2, Red-edge3,NIR	Sentinel-2
4	Normalized difference vegetation index	(NIR - Red) / (NIR+Red)	Sentinel-2
4	Modified Normalized Difference Water Index	(Green – SWIR1) / (Green + SWIR1)	Sentinel-2
4	Normalized Difference Built-up Index	(SWIR1- NIR) / (SWIR1+ NIR)	Sentinel-2
4	Enhanced Vegetation Index	2.5 * ((NIR - Red) / (NIR + 6 * Red - 7.5 * Blue + 1))	Sentinel-2
4	Burn Severity Index	((SWIR1 + Red) - (NIR + Blue)) / ((SWIR1 + Red) + (NIR + Blue))	Sentinel-2
4	Sentinel-2 Red-Edge Position Index	705 + 35 * (((RE3 + Red) / 2 - RE1) / (RE2 - RE1))	Sentinel-2
4	Green Normalized Difference Vegetation Index	(NIR - Green) / (NIR + Green)	Sentinel-2
4	Meris Terrestrial Chlorophyll Index	(RE2 - RE1) / (RE1 - Red)	Sentinel-2
32	NIR_textural features	Mean,Variance,Homogeneity,Contrast,Dissimilarity, Entroy,Angular Second Moment,Correlation	Sentinel-2

3	topographic data	DEM,slope,aspect	SRTM DEM

Table 1: Detailed description of all the features.

3. Methodology

3.1 workflow:

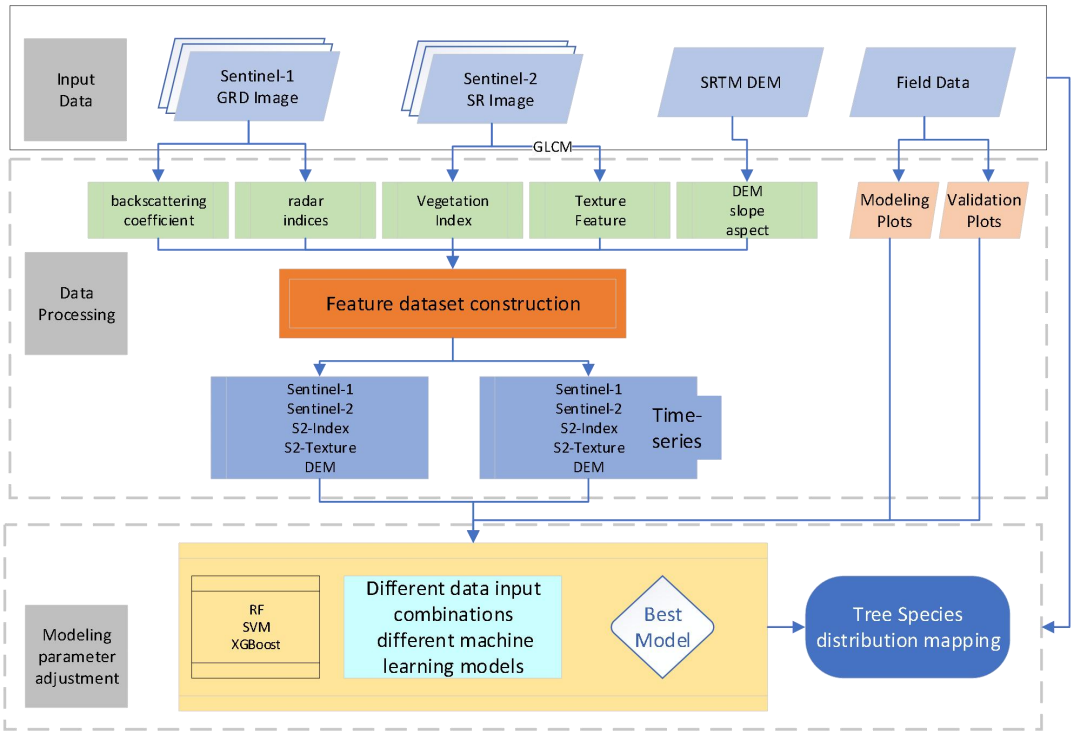

Figure 2: Workflow overview.

3.2 Machine Learning Methods:

The Random Forest algorithm, introduced by Breiman et al., is an ensemble learning method that leverages a collection of decision trees, specifically Classification and Regression Trees (CART), as base learners. In this approach, multiple decision trees independently vote on the classification of data, and the majority vote determines the final classification outcome of the Random Forest. This ensemble method significantly enhances classification accuracy compared to individual decision trees by reducing variance and improving generalization(Immitzer et al., 2012).

The XGBoost algorithm, known for its high efficiency in both execution speed and prediction accuracy, incorporates regularization terms in its cost function to control model complexity and mitigate overfitting. These regularization terms include penalties on leaf node weights and tree depth, contributing to the model's robustness and predictive performance(Zhou et al., 2022).

In this study, we employ both Random Forest and XGBoost algorithms to evaluate various feature combinations and determine the configurations that yield the highest classification

accuracy. We examine whether the integration of multi-temporal remote sensing images improves classification performance. For the Random Forest model, we optimize the `n_estimators` parameter to identify the most effective ensemble size for achieving the best classification results. In the case of the XGBoost model, we perform a grid search to fine-tune four key parameters—`n_estimators`, `learning_rate`, `max_depth`, and `subsample`—to determine the optimal parameter set for maximizing accuracy. This process aims to enhance the classification performance of dominant tree species by leveraging the combined strengths of these advanced machine learning techniques(Wongchai et al., 2022).

3.3 Accuracy Validation:

In Random Forest classification, the Out-of-Bag (OOB) error estimation is an intrinsic evaluation technique that utilizes Out-of-Bag samples—those not included in the bootstrap sample of a given tree—for assessing the model's generalization performance(Richter et al., 2016). This approach obviates the need for additional validation sets or cross-validation by leveraging the data inherently set aside during training. In our study, we perform OOB scoring by varying the `n_estimators` parameter, identifying the configuration with the highest OOB scores as the optimal model. These optimal models are then employed for subsequent classification tasks using the specified bands.

For the XGBoost model, K-fold cross-validation is employed as the parameter optimization method to mitigate the risk of overfitting and ensure robust model performance. During parameter optimization, the K-fold cross-validation technique is applied iteratively across different parameter settings to evaluate model accuracy(Pal & Patel, 2020). The average accuracy from ten cross-validation folds is used to determine the optimal parameters. The final model is configured using the optimal values for `n_estimators`, `learning_rate`, `max_depth`, and `subsample`, as derived from this parameter optimization process.

Following the selection of the optimal model and input feature combination, the final model's performance is evaluated using a confusion matrix(Persson et al., 2018). This matrix provides a detailed visualization of the classification performance by displaying actual versus predicted categories, with rows representing actual classes and columns representing predicted classes. Each cell in the matrix reflects the number of samples classified into each category pair. The confusion matrix enables the computation of various performance metrics, such as accuracy, precision, and recall, thereby offering a comprehensive assessment of the model’s performance across different classes. In this study, accuracy and recall rates are specifically utilized to evaluate and refine the model, with the final assessment being based on the detailed metrics provided by the confusion matrix.

4. Results

4.1 Input data combination and filtering

The input data are categorized into several distinct groups based on their characteristics. The Elevation Data Group includes topographic features such as Digital Elevation Model (DEM), slope, and aspect. The Sentinel-1 Data Group encompasses polarization bands and derived indices, including backscatter difference, backscatter division, and backscatter amplitude. The Sentinel-2 Data Group consists of raw band reflectance values, while the Sentinel-2 Index Group includes various vegetation indices such as the Normalized Difference Vegetation Index (NDVI), Modified Normalized Difference Water Index (MNDWI), Normalized Difference Built-up Index (NDBI), Enhanced Vegetation Index (EVI), Burn Severity Index (BSI), Sentinel-2 Red-Edge Position Index (S2REP), Green Normalized Difference Vegetation Index (GNDVI), and Meris Terrestrial Chlorophyll Index (MTCI). Additionally, the Sentinel-2 Texture Feature Group comprises texture features extracted from Sentinel-2's red-edge bands using the Gray-Level Co-occurrence Matrix (GLCM), including metrics such as mean, variance, homogeneity, contrast, dissimilarity, entropy, angular second moment, and correlation.These data are further organized according to their temporal characteristics, leading to the formation of various combinations. This approach integrates features from different periods to capture temporal variations and enhance the accuracy of tree species classification.

Name	Input data combination
s1_s2	Sentinel-1 group, Sentinel-2 group
s1_s2index	Sentinel-1 group, Sentinel-2 index group
s1_s2texture	Sentinel-1 group,Sentinel-2 texture feature group
s1_s2index_texture	Sentinel-1 group, Sentinel-2 index group ,Sentinel-2 texture feature group
s2_s2index	Sentinel-2 group, Sentinel-2 index group
s2_s2texture	Sentinel-2 group, Sentinel-2 texture feature group
s2index_s2texture	Sentinel-2 index group,Sentinel-2 texture feature group
all	Sentinel-1 group, Sentinel-2 group,Sentinel-2 index group, Sentinel-2 texture feature group
time_s1_s2(3,6,9,12)	Sentinel-1 group, Sentinel-2 group
time_s1_s2index(3,6,9,12)	Sentinel-1 group, Sentinel-2 index group
time_s1_s2texture(3,6,9,12)	Sentinel-1 group,Sentinel-2 texture feature group
time_s1_s2index_texture(3,6,9,12)	Sentinel-1 group, Sentinel-2 index group ,Sentinel-2 texture feature group
time_s2_s2index(3,6,9,12)	Sentinel-2 group, Sentinel-2 index group

time_s2_s2texture(3,6,9,12)	Sentinel-2 group, Sentinel-2 texture feature group
time_s2index_s2texture(3,6,9,12)	Sentinel-2 index group,Sentinel-2 texture feature group
time_all(3,6,9,12)	Sentinel-1 group, Sentinel-2 group,Sentinel-2 index group, Sentinel-2 texture feature group

Table 2: Data entry combination details.

The data combinations described above were utilized in both the Random Forest and XGBoost models. For the Random Forest model, parameter tuning was conducted for each data combination by varying the `n_estimators` parameter in increments of 5, ranging from 30 to 300. The model's performance was evaluated using the Out-of-Bag (OOB) error estimation method. The optimal parameter settings for each input data combination were determined based on accuracy metrics, revealing that the highest prediction accuracy and kappa coefficient were achieved when incorporating time series data from Sentinel-1, Sentinel-2 indices, and Sentinel-2 texture features.

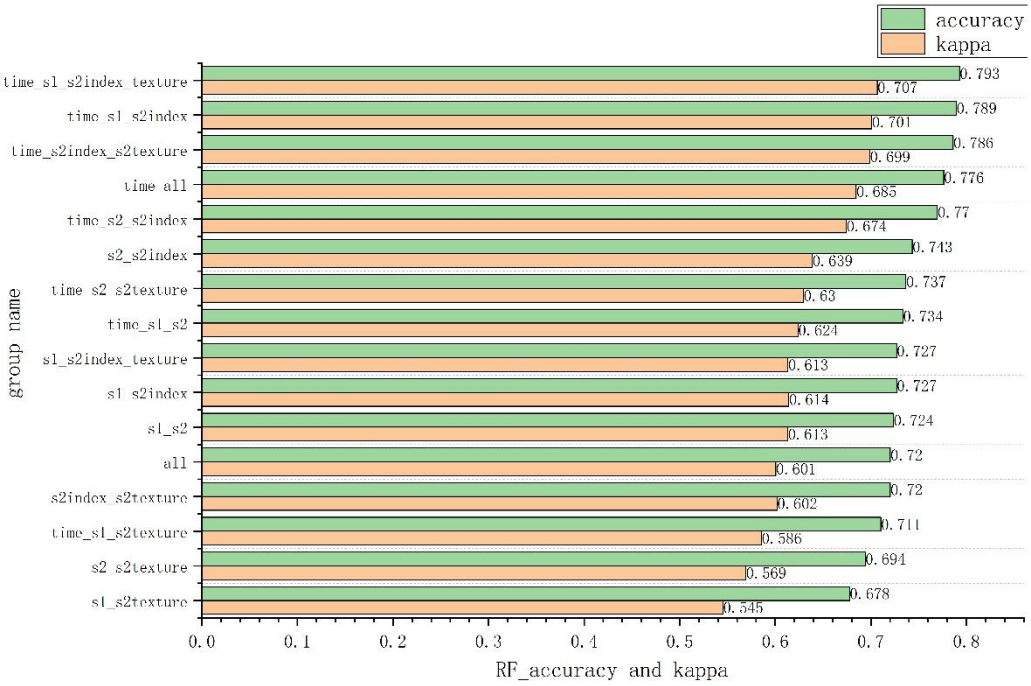

Figure 3: Random forest input data combination screening.

Similarly, in the XGBoost model, the aforementioned 16 data inputs were employed, with adjustments made to four key parameters: `n_estimators`, `learning_rate`, `max_depth`, and `subsample`. Through parameter optimization, it was found that the combination of Sentinel-1, Sentinel-2, Sentinel-2 indices, and Sentinel-2 texture features—encompassing all bands—yielded the highest prediction accuracy and kappa coefficient. Consequently, the final models

for subsequent predictions were based on these optimal data input combinations, which demonstrated superior performance in classification tasks.

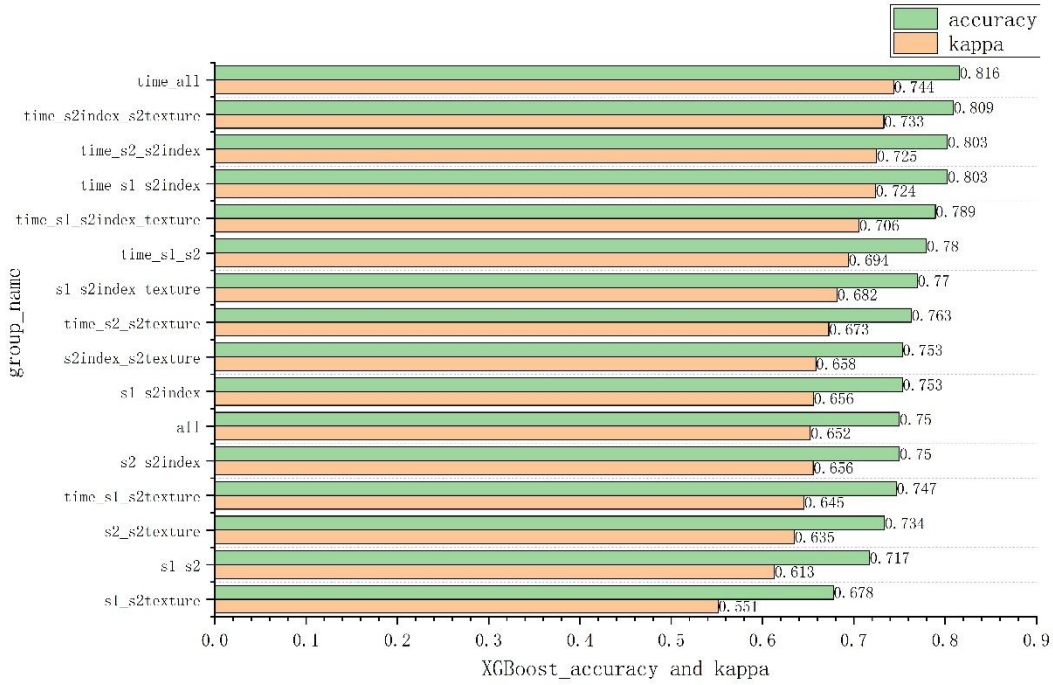

Figure 4: XGBoost input data combination screening.

Further parameter optimization was conducted for the selected data combinations in both the Random Forest and XGBoost models to achieve finer parameter tuning. For the Random Forest model, the `n_estimators` parameter was refined with a range from 30 to 500 in increments of 1. The results, as illustrated in the corresponding figure, indicated that the optimal parameter was 'n_estimators'= 371. This value was then utilized to construct the final Random Forest model.

In the case of the XGBoost model, a more detailed parameter tuning process was employed. Initially, only the `n_estimators` parameter was adjusted, with other parameters set to their default values. Following this, the parameters `max_depth` and `learning_rate` were optimized sequentially. The final optimal parameter settings were found to be `n_estimators` = 124, `learning_rate` = 0.1, and `max_depth` = 14. The model was constructed using these optimal parameters, and its accuracy was subsequently evaluated to ensure the effectiveness of the refined settings.

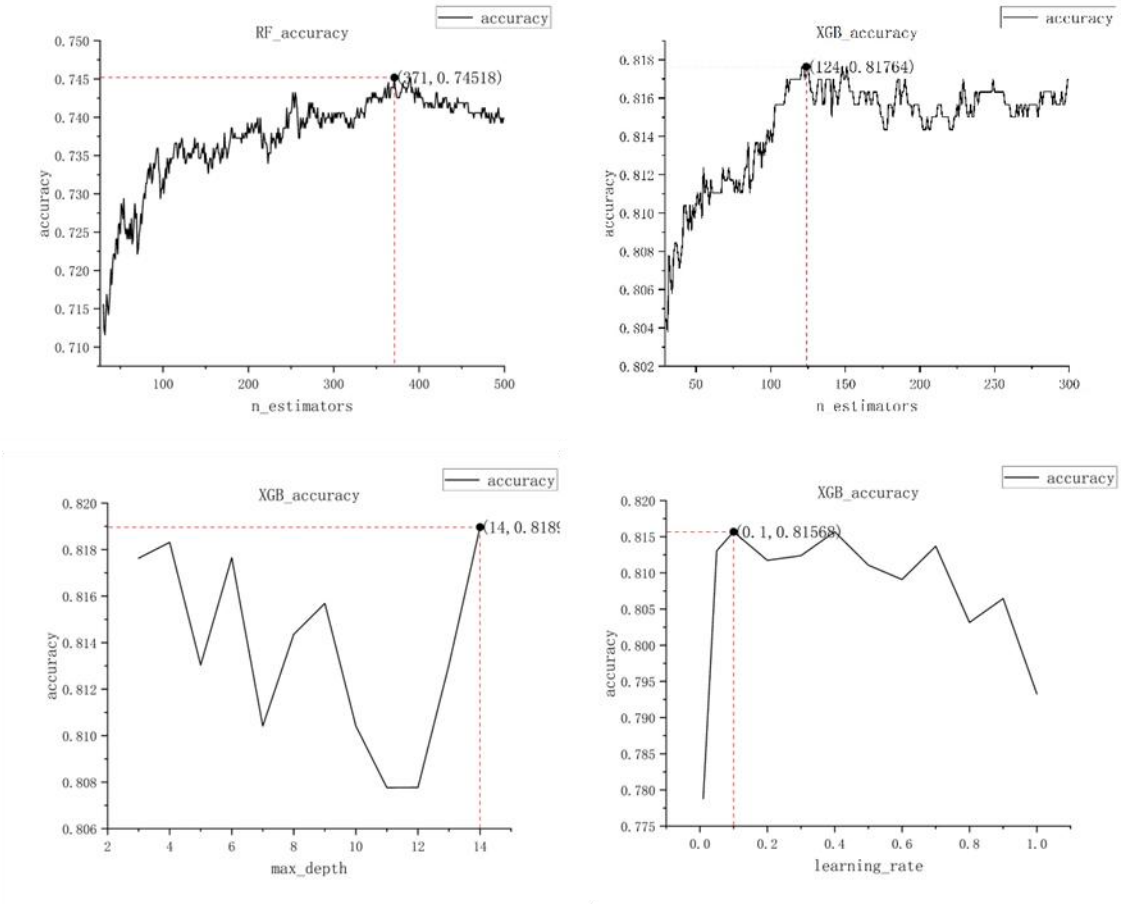

Figure 5: Curve of parameter adjustment between random forest and XGBoost model.

4.2 Feature importance

Upon finalizing the optimal models, feature importance was assessed and ranked to identify the most influential variables for classification. The results, detailed in the figure below, highlight the top 20 features based on their importance scores.

In the Random Forest (RF) model, the most significant features included the Normalized Difference Vegetation Index (NDVI) for March and the Enhanced Vegetation Index (EVI) for June. These were followed by the mean texture feature and the Green Normalized Difference Vegetation Index (GNDVI) for June. Additionally, elevation data emerged as a prominent feature, underscoring the substantial role of Sentinel-2 vegetation indices in the RF model. The texture features, while influential, were less dominant compared to the vegetation indices and elevation data.

In the XGBoost model, NDVI for March also exhibited the highest feature importance, with EVI and the mean texture index for June following closely. Sentinel-2 band data was similarly recognized as highly significant, reflecting the model's reliance on both vegetation indices and texture features for accurate classification.

These findings underscore the critical role of specific vegetation indices and topographic features in enhancing the predictive performance of both Random Forest and XGBoost models.

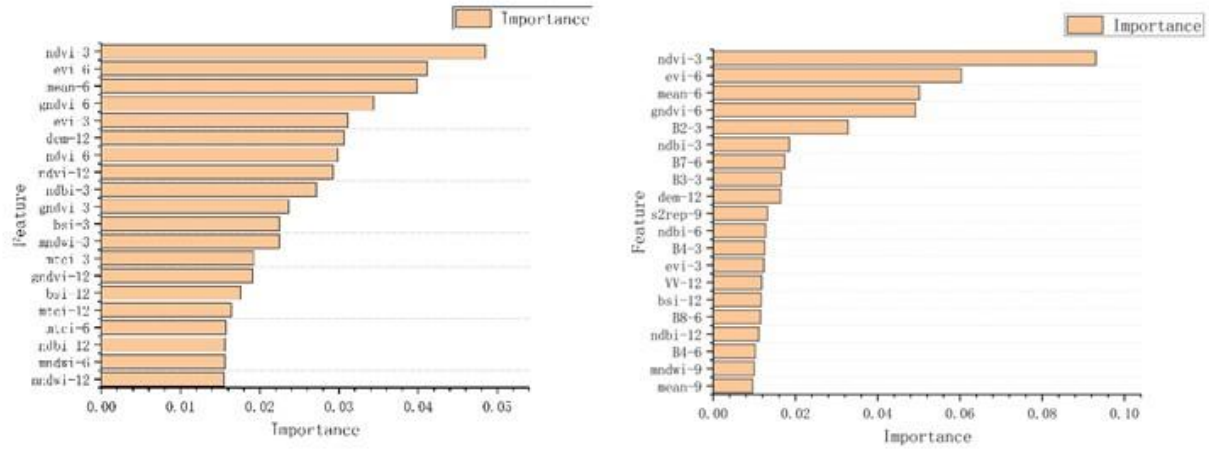

Figure 5: Feature importance ranking results of random forest and XGBoost models.

4.3 Precision comparison and mapping of tree species classification results

Through the selection of the optimal data combinations and the construction of the most suitable models, the establishment of the classification models was successfully completed, yielding accuracy evaluation results for both the Random Forest and XGBoost models. The performance of each model was assessed using key metrics such as overall accuracy and the kappa coefficient, ensuring a comprehensive evaluation of their classification capabilities. These results underscore the effectiveness of the selected features and the applied methodologies in enhancing tree species classification accuracy across the study area.

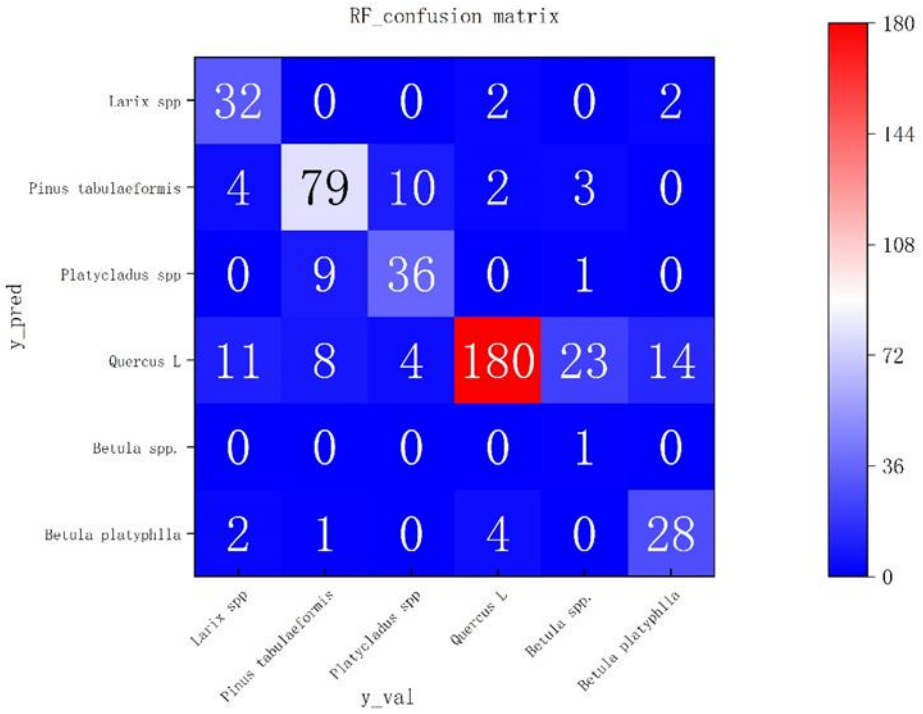

Figure 6: The confusion matrix of six target tree species of random forest model.

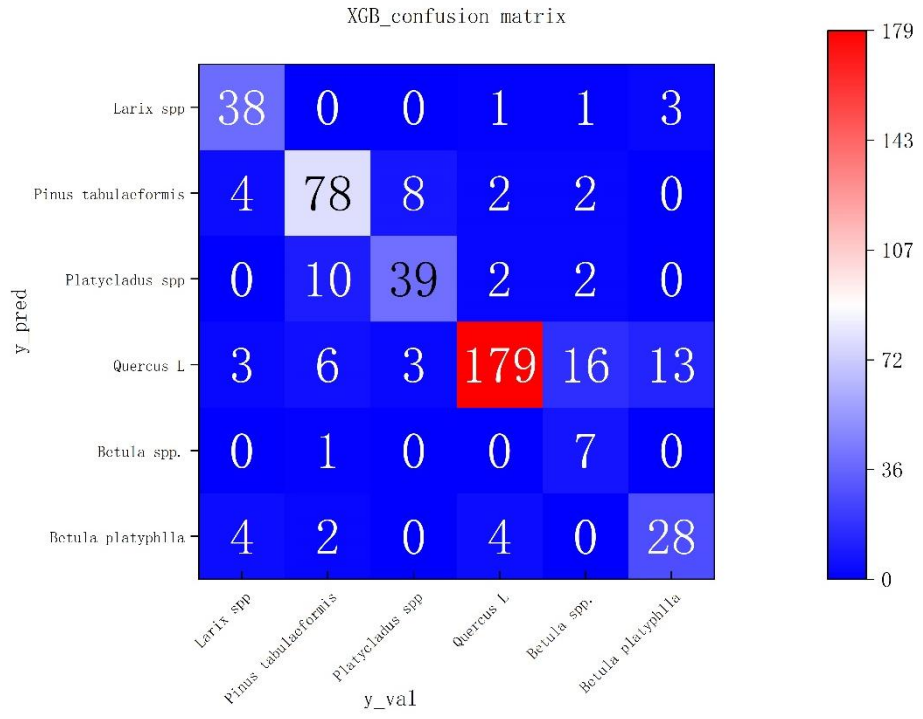

Figure 7: The confusion matrix of six target tree species of XGBoost model.

Based on the classification results, the Random Forest model achieved an accuracy of 78.07% with a kappa coefficient of 0.691, while the XGBoost model demonstrated superior performance with an accuracy of 81.25% and a kappa coefficient of 0.737. A comparison of

the two models indicates that the XGBoost model outperformed the Random Forest model in both accuracy and kappa coefficient, and thus the XGBoost model was selected for further application in the final mapping process.

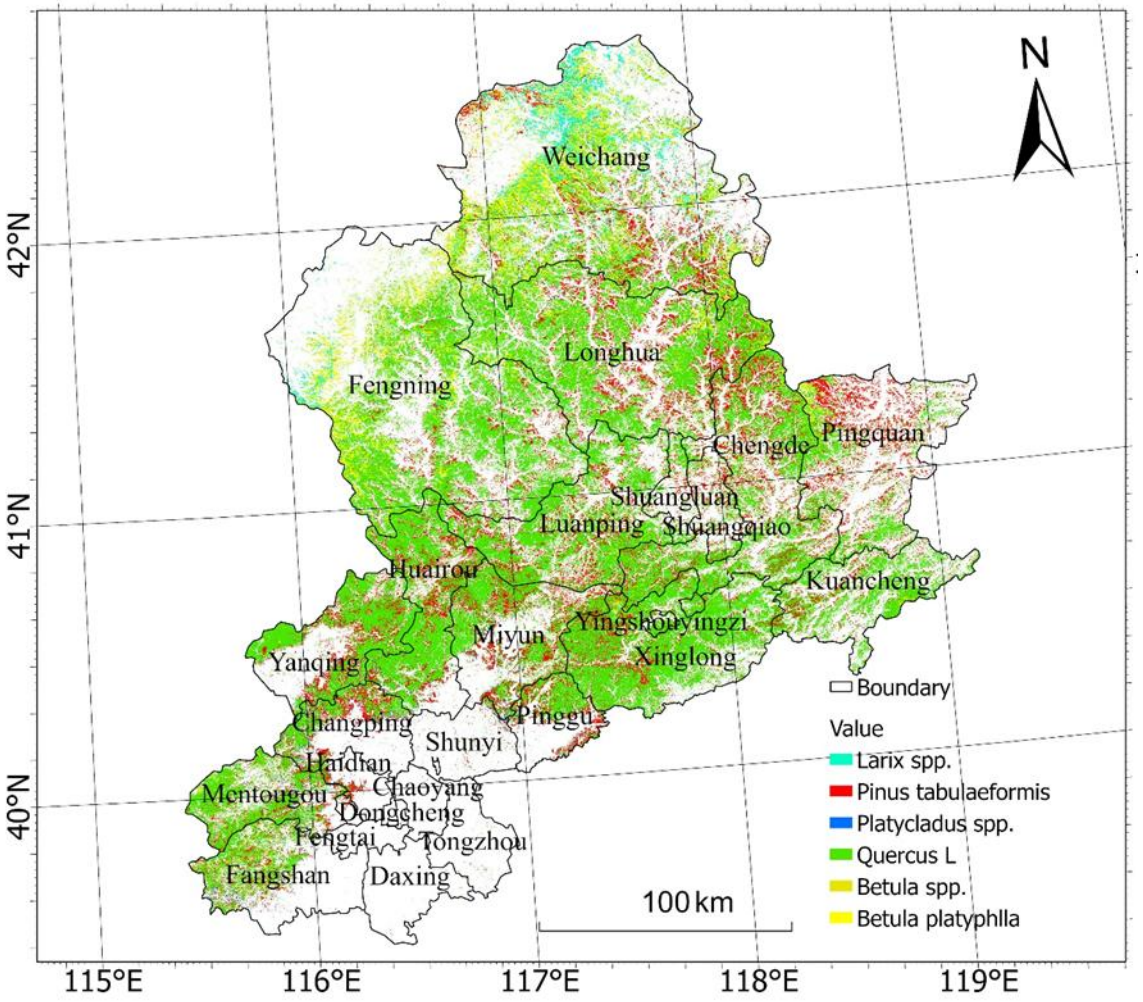

Figure 8: Distribution mapping of the six dominant tree species.

Utilizing the XGBoost model, the tree species distribution map for the study area was generated based on the processed remote sensing data. Furthermore, a statistical analysis of the resulting tree species distribution was conducted, and a pie chart summarizing the area distribution of the dominant species was produced. The analysis reveals that species such as Quercus mongolica dominate the study area, covering approximately 66% of the total area, with extensive distribution throughout the region. Larix spp. is primarily found in the northern part of Chengde City, particularly in Weichang County. Pinus tabuliformis represents the second most widespread species, with a distribution spanning the entire study area. Platycladus orientalis is predominantly located in the Beijing region, while Betula spp.

is mainly concentrated in northern Chengde. The specific distribution and area proportions of each tree species are indicated in the accompanying figure.

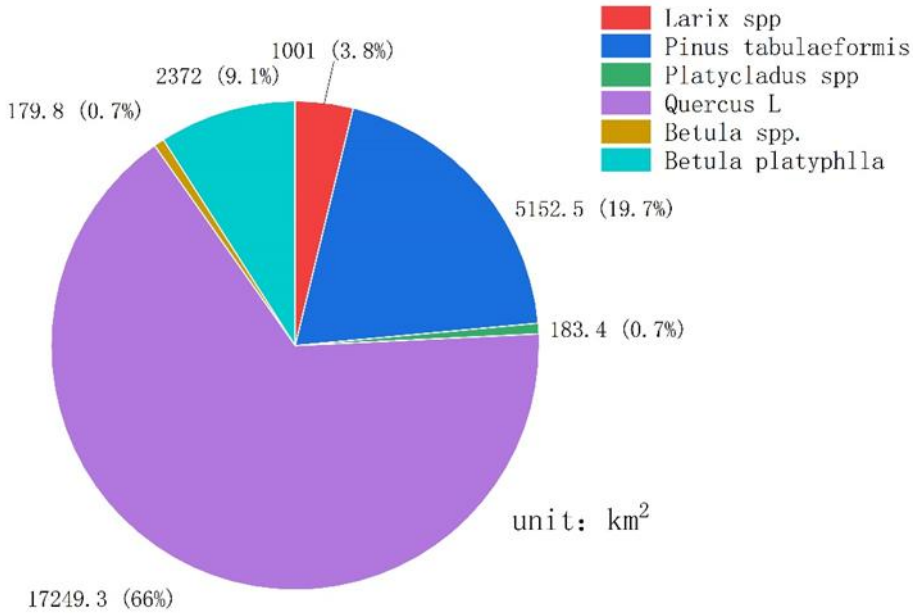

Figure 9: Statistical mapping of the area of the six dominant tree species.

5. Discussion

Currently, numerous methods exist for the classification of forest tree species with high accuracy; however, most research focuses on areas such as forest farms and parks where tree species are uniform and canopy coverage is high. These regions provide easier access to UAV LiDAR or hyperspectral data, and many studies leverage such data to employ deep learning techniques for image classification. Nevertheless, the findings from these studies are difficult to generalize to larger regions, making it challenging to achieve large-scale tree species classification and mapping. As a result, these methods are often unsuitable for applications such as national forest resource inventories, environmental monitoring, or carbon cycle research. Therefore, methods based on multi-spectral data fusion offer significant advantages, and there is a clear need to further explore multi-spectral approaches for tree species classification.

In this study, Random Forest (RF) and XGBoost models were compared for tree species classification, with both models yielding high classification accuracy. The XGBoost model slightly outperformed the Random Forest model in this study area. Additionally, the results demonstrated that classification accuracy improves as more input features are incorporated. Specifically, the use of multi-temporal indices significantly enhances the models' performance. Generally, adding more features leads to higher classification accuracy,

suggesting that future research should explore the inclusion of additional indices that reflect differences between tree species to further improve model performance.

Among the input features, various vegetation indices calculated from Sentinel-2 data exhibited high importance, indicating that different tree species show significant variability in visible light indices. Moreover, texture features, such as the mean texture feature, also ranked highly in importance. For Sentinel-1 data, backscatter coefficients were found to be more important than derived indices. Regarding Sentinel-2 reflectance data, the March and June reflectance bands proved to be the most important, reflecting that during the autumn and winter months, the reflectance differences between tree species diminish as most forests in the study area consist of deciduous broad-leaved species. During these months, the primary reflectance content shifts from vegetation to soil, reducing the importance of reflectance features in classification, a trend also observed in the vegetation indices.

According to the classification results, Quercus L and other oaks were the dominant tree species in Chengde City, with Pinus tabuliformis being the second most prevalent species. These two species are widely distributed across the study area. In contrast, Platycladus spp. was predominantly distributed in Beijing, while Betula spp. and Larix spp. were mainly found in Chengde. Despite the proximity of Beijing and Chengde, their different altitudes result in distinct species distributions, with Betula and Larix species more common in Chengde, and Platycladus prevalent in Beijing.

6. Conclusion

The primary objective of this study is to classify the dominant tree species across a large forested region. This paper presents a novel approach for identifying and classifying tree species in Chengde City and Beijing using a combination of Sentinel-1 and Sentinel-2 satellite data, incorporating multiple data input combinations and temporal datasets. As a result, a 10-meter resolution classification map of the dominant tree species was generated. The tree species classification data obtained through this study can provide essential decision-making support for government agencies involved in forest management, planting, and monitoring.

Through the processing of extensive datasets, six major tree species were classified within the regions of Chengde and Beijing, including larch, tabaric pine, cypress, oak, and birch. The results indicated that oak species, such as Quercus L, are the predominant tree species, covering 66% of the study area. Larch occupies 19.7% of the area, while birch and cypress cover the smallest proportions.

By utilizing freely available Sentinel-1 and Sentinel-2 data, this study successfully generated a large-scale forest tree species distribution map. Validation with ground-truth data confirmed the reliability of these satellite data for tree species classification, demonstrating their significant potential for broader applications in forest resource monitoring. Moreover, the results underscore the effectiveness of machine learning algorithms, such as Random Forest and XGBoost, in achieving accurate tree species classification across extensive forested landscapes.

References

Abdollahnejad, A., & Panagiotidis, D. (2020). Tree species classification and health status assessment for a mixed broadleaf-conifer forest with UAS multispectral imaging. *Remote Sensing, 12*(22), 3722.

Alexander, C., Tansey, K., Kaduk, J., Holland, D., & Tate, N. J. (2010). Backscatter coefficient as an attribute for the classification of full-waveform airborne laser scanning data in urban areas. *ISPRS journal of photogrammetry and remote sensing, 65*(5), 423-432.

Arvor, D., Jonathan, M., Meirelles, M. S. P., Dubreuil, V., & Durieux, L. (2011). Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. *International Journal of Remote Sensing, 32*(22), 7847-7871.

Bhattarai, R., Rahimzadeh-Bajgiran, P., Weiskittel, A., Homayouni, S., Gara, T. W., & Hanavan, R. P. (2022). Estimating species-specific leaf area index and basal area using optical and SAR remote sensing data in Acadian mixed spruce-fir forests, USA. *International Journal of Applied Earth Observation and Geoinformation, 108*, 102727.

Choi, S., Lee, W.-K., Kwak, D.-A., Lee, S., Son, Y., Lim, J.-H., & Saborowski, J. (2011). Predicting forest cover changes in future climate using hydrological and thermal indices in South Korea. *Climate Research, 49*(3), 229-245.

Coluzzi, R., Imbrenda, V., Lanfredi, M., & Simoniello, T. (2018). A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses. *Remote Sensing of Environment, 217*, 426-443.

Cornelis, J.-T., Ranger, J., Iserentant, A., & Delvaux, B. (2010). Tree species impact the terrestrial cycle of silicon through various uptakes. *Biogeochemistry, 97*, 231-245.

Dalponte, M., Ørka, H. O., Ene, L. T., Gobakken, T., & Næsset, E. (2014). Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data. *Remote Sensing of Environment, 140*, 306-317.

Dalponte, M., Ørka, H. O., Gobakken, T., Gianelle, D., & Næsset, E. (2012). Tree species classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing, 51*(5), 2632-2645.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., & Martimort, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment, 120*, 25-36.

Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., Straub, C., & Ghosh, A. (2016). Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment, 186*, 64-87.

Franklin, S. E., & Ahmed, O. S. (2018). Deciduous tree species classification using object-based analysis and machine learning with unmanned aerial vehicle multispectral data. *International Journal of Remote Sensing, 39*(15-16), 5236-5245.

Fujimoto, A., Haga, C., Matsui, T., Machimura, T., Hayashi, K., Sugita, S., & Takagi, H. (2019). An end to end process development for UAV-SfM based forest monitoring: Individual tree detection, species classification and carbon dynamics simulation. *Forests, 10*(8), 680.

Immitzer, M., Atzberger, C., & Koukal, T. (2012). Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data. *Remote Sensing, 4*(9), 2661-2693.

Li, D., Ke, Y., Gong, H., & Li, X. (2015). Object-based urban tree species classification using bi-temporal WorldView-2 and WorldView-3 images. *Remote Sensing, 7*(12), 16917-16937.

Madonsela, S., Cho, M. A., Ramoelo, A., Mutanga, O., & Naidoo, L. (2018). Estimating tree species diversity in the savannah using NDVI and woody canopy cover. *International Journal of Applied Earth Observation and Geoinformation, 66*, 106-115.

Majasalmi, T., Eisner, S., Astrup, R., Fridman, J., & Bright, R. M. (2018). An enhanced forest classification scheme for modeling vegetation–climate interactions based on national forest inventory data. *Biogeosciences, 15*(2), 399-412.

Michałowska, M., & Rapiński, J. (2021). A review of tree species classification based on airborne LiDAR data and applied classifiers. *Remote Sensing, 13*(3), 353.

Ming, S., Xiong, L., Zhicheng, L., Yunlu, Z., & Chaonan, C. (2021). Evolution analysis and optimization research of ecosystem service value in Chengde City, Hebei Province of northern China based on land use/land cover change (LUCC). *Journal of Beijing Forestry University, 43*(3), 106-116.

Onojeghuo, A. O., Onojeghuo, A. R., Cotton, M., Potter, J., & Jones, B. (2021). Wetland mapping with multi-temporal sentinel-1 &-2 imagery (2017–2020) and LiDAR data in the grassland natural region of alberta. *GIScience & Remote Sensing, 58*(7), 999-1021.

Otsu, K., Pla, M., Duane, A., Cardil, A., & Brotons, L. (2019). Estimating the threshold of detection on tree crown defoliation using vegetation indices from UAS multispectral imagery. *Drones, 3*(4), 80.

Pal, K., & Patel, B. V. (2020). Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. 2020 fourth international conference on computing methodologies and communication (ICCMC),

Persson, M., Lindberg, E., & Reese, H. (2018). Tree species classification with multi-temporal Sentinel-2 data. *Remote Sensing, 10*(11), 1794.

Richter, R., Reu, B., Wirth, C., Doktor, D., & Vohland, M. (2016). The use of airborne hyperspectral data for tree species classification in a species-rich Central European forest area. *International Journal of Applied Earth Observation and Geoinformation, 52*, 464-474.

Sabins Jr, F. F., & Ellis, J. M. (2020). *Remote sensing: Principles, interpretation, and applications*. Waveland Press.

Sesnie, S. E., Gessler, P. E., Finegan, B., & Thessler, S. (2008). Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment, 112*(5), 2145-2159.

Sun, C., Li, J., Liu, Y., Liu, Y., & Liu, R. (2021). Plant species classification in salt marshes using phenological parameters derived from Sentinel-2 pixel-differential time-series. *Remote Sensing of Environment, 256*, 112320.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., & Brown, M. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment, 120*, 9-24.

Wongchai, W., Onsree, T., Sukkam, N., Promwungkwa, A., & Tippayawong, N. (2022). Machine learning models for estimating above ground biomass of fast growing trees. *Expert Systems with Applications, 199*, 117186.

Yang, G., Zhao, Y., Li, B., Ma, Y., Li, R., Jing, J., & Dian, Y. (2019). Tree species classification by employing multiple features acquired from integrated sensors. *Journal of Sensors, 2019*(1), 3247946.

Zhou, G., Ni, Z., Zhao, Y., & Luan, J. (2022). Identification of bamboo species based on Extreme Gradient Boosting (XGBoost) using Zhuhai-1 orbita hyperspectral remote sensing imagery. *Sensors, 22*(14), 5434.