

A Sea Ice Recognition Algorithm in Bohai Based on Random Forest

Tao Li¹, Di Wu¹, Rui Han², Jinyue Xia³ and Yongjun Ren^{4,*}

¹School of Artificial Intelligence/School of Future Technology, Nanjing University of Information Science and Technology, Nanjing, 210044, China

²Unit 93117 of PLA, PLA, Jiangsu, 210000, China

³International Business Machines Corporation (IBM), NY, 100014, USA

⁴School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 210044, China

*Corresponding Author: Yongjun Ren. Email: renyj100@126.com

Received: 08 March 2022; Accepted: 07 May 2022

Abstract: As an important maritime hub, Bohai Sea Bay provides great convenience for shipping and suffers from sea ice disasters of different severity every winter, which greatly affects the socio-economic and development of the region. Therefore, this paper uses FY-4A (a weather satellite) data to study sea ice in the Bohai Sea. After processing the data for land removal and cloud detection, it combines multi-channel threshold method and adaptive threshold algorithm to realize the recognition of Bohai Sea ice under clear sky conditions. The random forests classification algorithm is introduced in sea ice identification, which can achieve a certain effect of sea ice classification recognition under cloud cover. Under non-clear sky conditions, the results of Bohai Sea ice identification based on random forests have been improved, and the algorithm can effectively identify Bohai Sea Ice and can improve the accuracy of sea ice identification, which lays a foundation for the accuracy and stability of sea ice identification. It realizes sea ice identification in the Bohai Sea and provides data support and algorithm support for marine climate forecasting related departments.

Keywords: FY-4A; random forests; Bohai Sea Ice; sea ice identification; adaptive threshold

1 Introduction

As one of the more sensitive environmental factors in the climate environment, sea ice has an important impact on the balance of materials and heat exchange in the region [1–3]. With the development and progress of social life, people have become more and more aware of sea ice, which has a wide range of impacts on shipping, construction of port facilities, offshore oil extraction, etc., and poses a threat to sailing vessels. Sea ice disasters have a great impact, and people must monitor and forecast them effectively.

Bohai Sea waters are in a relatively special geographical location, the climate conditions are relatively unique, and the region in the winter of each year will suffer a certain amount of sea



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ice disaster. Bohai Sea Ice is seasonal sea ice, every winter there will be different degrees of icing phenomenon, and this icing phenomenon is affected by cold air, icing period is generally December of the first year to March of the second year, lasting about For 3–4 months, seriously affecting the production operations in the surrounding areas [4]. As one of the major marine disasters, sea ice in the Bohai Sea can seriously affect socio-economic and production activities. When ice conditions are severe, the sea ice area can even cover more than 70% of the entire Bohai Sea area, which has a huge impact on normal shipping, infrastructure construction and oil and gas extraction work in the Bohai Sea area. Therefore, accurate monitoring and forecasting of sea ice in the Bohai Sea is particularly important. Fraser proposed an algorithm to remove cloud cover in polar regions by synthesizing MODIS (Moderate-resolution Imaging Spectroradiometer) satellite images and generated images of sea ice around the Mertz Glacier region in East Antarctica [5]. Combining the cloud spectrum and reflectivity of MODIS cloud mask products, Riggs improves the cloud mask algorithm, reduces the occlusion of snow and ice by clouds, and improves the ability to draw snow scenes [6].

The research on sea ice began in the 1850s. In 1960, the United States successfully launched the first weather satellite TIROS-1 (weather satellite) satellite, marking that satellite remote sensing technology began to become an important means of obtaining ocean information [7]. The first NOAA/AVHRR (The third generation of meteorological observation satellites/Advanced Very High Resolution Radiometer) satellite was successfully launched in 1978, and by 1986 the satellite was widely used in sea ice monitoring [8]. Key et al. pointed out that anomalies occur when AVHRR (Advanced Very High Resolution Radiometer) data are used in cloud-covered seas [9]. Lindsay et al. used AVHRR visible light and near-infrared data to identify sea ice density [10]. Spinhime et al. successfully achieved sea ice detection on the East Antarctic coast using MODIS data, but the detection results are still affected by cloud cover [11]. JJ Yackel et al. estimated the coverage of summer surface melt zone of arctic sea ice based on MODIS data [12]. M Mäkynen et al. used MODIS data to identify sea ice thickness in the Barents and Kara Seas by combining reflectance and nighttime ice surface temperature [13].

Remote sensing is acquiring details about an object without any physical interaction with that object [14]. The rapid development of satellite remote sensing technology has provided the possibility of acquiring continuous spatial and temporal data for sea ice monitoring and forecasting [15]. Satellite-based sea ice remote sensing has the advantages of easy data acquisition, wide data coverage and data collection without special geographic environment restrictions. The use of satellite remote sensing data can achieve effective and more accurate identification of sea ice in the Bohai Sea, and remote sensing technology is increasingly applied in the field of sea ice disaster monitoring [16]. With the advances in sensor technology and availability of high resolution satellite imagery, monitoring of changes on earth's surface has become easier [17].

As a new generation of Chinese geostationary meteorological satellite, FY-4A carries the Advanced Geosynchronous Radiation Imager (AGRI), which has 14 observation channels. At present, satellite remote sensing data are widely used in sea ice identification, but the research on sea ice identification using FY-4A data is still in the early stage. Moreover, due to the defects of satellite remote sensing data itself, sea ice identification only has better effect under clear sky conditions, and under non-clear sky conditions, sea ice is influenced by clouds, so it cannot get better identification effect, and the availability of data is greatly reduced. In the context of the rapid development of machine learning-related algorithms and their wide application in various fields, random forests algorithm has been widely used in satellite remote sensing image classification and has good application effect [18]. This paper uses the high time resolution characteristics of FY-4A satellite data, combined with sea ice business data, and uses random forests classification algorithm to study the Bohai Sea ice

identification, and establishes a Bohai Sea ice identification model based on the random forest algorithm, which is used in the Bohai Sea area. Monitoring of sea ice.

2 Relevant Methods

2.1 Ensemble Learning

Ensemble learning [19] is one of the hot research directions in the current artificial intelligence field. It is a kind of algorithm that combines multiple learners through an ensemble method or strategy to obtain a better performance and a more comprehensive algorithm. This is a learning method that draws on the best of others. The key is that each base learner has a certain accuracy, and there are also differences. The random forests method is representative of ensemble learning algorithms. These are algorithms based on decision trees and are widely used by many data researchers in various fields.

2.2 Random Forests

Random forests, PageRank (Page Sorting Algorithm), support vector machine, K-Means (A machine learning method) and other algorithms are called the top ten algorithms of machine learning [20]. Among several machine learning methods, random forests is a robust method [21]. Random Forests have strong generalization ability and is not easy to overfit [22]. The random forests algorithm, as a classification algorithm, is widely used because of its superior performance [23]. The Random Forests [24] method is representative of ensemble learning algorithms. These algorithms are based on decision trees and are widely used by many data researchers in various fields. Ensemble learning is one of the current hot research directions in artificial intelligence, which is a way to combine multiple learners by some integration method or strategy to obtain better and more comprehensive algorithms. The key to this is that each base learner has a certain degree of accuracy and also variability. Random Forests method is representative of integrated learning algorithms, and these are decision tree-based learners that are widely used by many data researchers in various fields. Khan uses random forests to predict the compressive strength of FA-based geopolymers concrete [25]. Jackins uses random forests for disease prediction of patients, which outperforms the naïve classifier [26]. Dawod uses random forests to classify mangroves and achieves higher values for precision, recall, F-score and overall accuracy [27].

3 Data

3.1 Study Area

The study area in this paper is the Bohai Sea, and the specific geographical location is 37°–42°N, 117°–123°E. The Bohai Sea is Chinese highest latitude sea area and one of the main areas where low temperatures exist in winter in China. It includes three bays, namely Bohai Bay, Liaodong Bay and Laizhou Bay. The geographical location of Bohai Sea is shown in Fig. 1. Liaodong Bay is the most severe area of sea ice among the three bays. The Bohai Sea is affected by cold air, geographic location, and complex submarine topography, making it subject to varying degrees of sea ice hazards each winter [28]. Usually, the Bohai Sea area from December will start to appear sea ice phenomenon, sea ice from the north to the south, from the coast to the sea deep water area trend development, icing period from December to March of the next year ranging, belong to the annual sea ice.

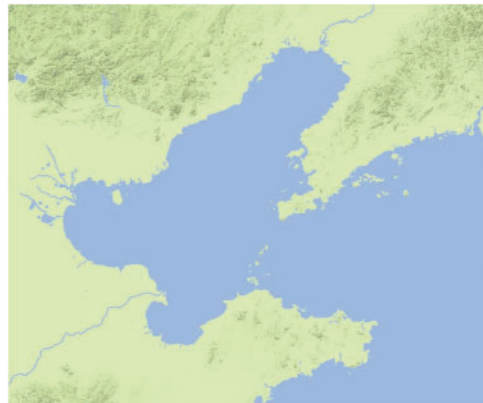


Figure 1: Study area

3.2 FY-4A

There are reflectance differences between ice and water in different spectral channels. The reflectance of seawater and snow decreases with increasing spectral channel wavelength, while the reflectance curve of ice has a single-peaked shape. In the visible light band, the reflectivity of snow is as high as 80%, while the reflectivity of seawater in this band is about 10% or less, and the reflectivity decreases as the wavelength increases. In the near-infrared wavelengths, the reflectivity of snow and ice is significantly lower than that of their respective visible wavelengths. However, there is still a big difference with the reflectivity of seawater in this band. This feature is convenient for the identification of sea ice, and other substances do not have this unique spectral feature.

FY-4A was successfully launched on Dec. 11, 2016, and officially received satellite data on May 8, 2018, with the satellite carrying a multi-channel scanning imaging radiometer for rapid area scanning at the minute level. The FY-4A satellite is Chinese second-generation geostationary orbit quantitative remote sensing meteorological satellite. It adopts three-axis stability control technology and can provide important support for weather forecasting and disaster warning in China and the Asia-Pacific region. FY-4A satellite covers visible, short-wave infrared, mid-wave infrared and long-wave infrared wavelengths, with 14 radiation imaging channels, close to the 16 channels of equivalent third-generation satellite products in Europe and the United States [29]. Tab. 1 shows FY-4A. Each band parameter of imaging radiometer data loaded by satellite. Compared with polar-orbiting satellites, FY-4A data has high time resolution. Among them, the average observation data of the full disk is 40 pieces per day, and the average observation data of China is 6 pieces of data per hour, and the average of 165 pieces of data per day. FY-4A data provides data support for the quasi-real-time monitoring of large-scale sea ice in the Bohai Sea.

Table 1: Parameters of each band of FY-4A AGRI data

Sensor	Waveband serial number	Center wavelength/ μm	Resolution/km
	1	0.47	1.0
	2	0.65	0.5

(Continued)

Table 1: Continued

Sensor	Waveband serial number	Center wavelength/ μm	Resolution/km
AGRI/FY-4A	3	0.825	1.0
	4	1.375	2.0
	5	1.61	2.0
	6	2.25	2.0
	7	3.75	2.0
	8	3.75	4.0
	9	6.25	4.0
	10	7.1	4.0
	11	8.5	4.0
	12	10.7	4.0
	13	12.0	4.0
	14	13.5	4.0

3.3 Land Removal

The land surface interferes with sea ice identification, and the spectral information of fixed ice areas covered by coastal waters in winter is so close to the land surface snow on shore that it can be easily confused. But land information is basically fixed, and the boundary between land and sea does not move with the seasons. In order to reduce the interference of land on sea ice recognition and retain the information of the Bohai Sea area, the land information needs to be removed first after data preprocessing. The principle of most land removal is to use the reflectivity characteristics of land and sea to distinguish by the waveband threshold method. This method needs to combine the actual situation of the study area to determine the threshold value, and the land and sea-land interface is more complex, and there are more classes of target interferents, leading to some errors in the results, Therefore the key issue is to perform sea-land separation efficiently and accurately, and to remove as much complex land information as possible by precise land removal.

Digital Terrain Model (DTM) digitally describes the ground terrain by various geomorphological factors such as slope, direction, and elevation [30]. Digital Elevation Model (DEM) is one of the subtypes of the digital terrain model, which is a digital simulation of ground terrain using only terrain elevation data, where the value of each grid is an elevation value.

For the problem of land disturbance, this paper uses Digital Elevation Model based land removal. The elevation data are provided by the National Geophysical Center (NGDC) of the United States. ETOPO1 (a topographic data) is a topographic elevation data for the whole world with a spatial resolution of 1 arc minute, which is a solid ground model that represents ground elevation information through a set of ordered arrays, providing basic information about the height of the Earth's surface and related features. Fig. 2 shows the results of land removal based on elevation data, where blue indicates ocean and yellow indicates land.



Figure 2: Land removal based on elevation data

3.4 Cloud Detection

Clouds are composed of water droplets and ice crystals in the atmosphere. If they are strongly absorbed in the infrared band, the infrared radiation below the cloud cannot be detected, but the atmospheric radiation above the cloud top can only be detected, resulting in the radiation value observed by the satellite and the actual value, Causes large deviations between the radiation values observed by satellites and the actual values [31]. Cloud detection refers to the estimation of total cloud volume, high cloud volume and cloud top height. The basic principle of cloud detection is to detect the difference in reflectivity and radiance values between clouds and soil, water, ice, snow and vegetation in the visible and infrared bands. Clouds usually have lower brightness temperature values and higher reflectance. In the visible light band, the reflectivity of the underlying surface depends on the difference in the types of ground objects. Among them, the reflectivity of vegetation is the lowest, followed by water, soil and towns have the highest reflectivity, and the reflectivity of clouds is significantly higher than that of the underlying surface, and Changes with height and thickness. In addition, the cloud top temperature is lower in the thermal infrared band, so the brightness temperature of the cloud in the remote sensing image is also lower [32].

In the process of sea ice identification, the Bohai Sea region is severely disturbed by clouds, which poses a great limitation to the use of optical remote sensing data such as FY-4A data. When using FY-4A visible data to identify sea ice in the Bohai Sea, it is easy to misjudge the sea ice due to the presence of clouds, so cloud detection is needed after land removal to effectively eliminate cloud interference.

The cloud detection in this article uses the real-time China cloud detection product of FY-4A satellite data. At present, the cloud detection method of FY-4A uses the standard deviation of the brightness temperature of the 12 μm channel and sets the appropriate characteristic threshold, combined with AGRI's The visible light channel image then obtains the real-time product of cloud detection [33]. Use cloud detection product data to obtain the result of cloud area recognition, and then generate binary cloud detection data, that is cloud mask.

4 Sea Ice Recognition

4.1 Multichannel Adaptive Thresholding Algorithm for Sea ice Recognition

Since the reflectivity and brightness temperature of ground objects will vary slightly with the seasons, there will be differences in remote sensing data parameters at different times and spaces. In addition, the data coverage in satellite remote sensing images is large, and it is difficult to achieve a good

segmentation effect if a fixed threshold is used. In this paper, based on the multi-channel threshold method, an adaptive threshold algorithm is used, which is applied to sea ice identification to improve the sea ice identification results.

Early object detection algorithms based on hand-crafted features mainly rely on the information of objects' edges and key feature points, and adopt the sliding window operation to determine the location of the object by scanning the whole image [34]. The threshold method is a region-based image segmentation algorithm, and its basic principle is to classify images by setting several threshold intervals. At present, the global empirical threshold segmentation method is widely used in the commonly used sea ice identification methods. During image processing and analysis, the adaptive thresholding algorithm can automatically adjust the processing method, constraints or processing parameters according to the data characteristics of the image target, so that these conditions are adapted to the statistical distribution characteristics and structural features of the image.

OTSU(Japanese scholar OTSU proposed in 1979) method [35] is an adaptive thresholding image segmentation method [36], which is based on the principle of least squares and is suitable for the automatic acquisition of global thresholds in the case of double peaks. The OTSU is used in order to calculate the threshold value to obtain the connected region and then use this threshold to binarize the study region.

In the field of image processing, the most common problem is to extract the target object from the image background. Theoretically, there are two distinct peaks on the image histogram and deep and sharp peaks and valleys between the two peaks, in which case the threshold should be set to the value of the peaks and valleys. In practice, however, the peaks and valleys between the bimodal peaks are not all deep and sharp, but flat and scattered, or the gap between the two peaks is too large, when the value of the peaks and valleys is uncertain. OTSU can solve such problems, specifically, assuming that the grayscale range of the image is $[0, L-1]$, the number of pixels is N , and the number of pixels with gray level i in the image is denoted by n_i , then $N = n_1 + n_2 + \dots + n_i$. where the probability of occurrence of a pixel with gray level i is:

$$P_i = \frac{n_i}{N}, P_i \geq 0, \sum_{i=0}^{L-1} P_i = 1 \quad (1)$$

Suppose the pixels in the image are divided into two classes O1 and O2 by thresholding t . O1 and O2 denote the target and background of the image, respectively. Where O1 is the set of all pixels with gray levels in the range of $[0, t]$ and O2 is the set of all pixels with gray levels in the range of $[t+1, L-1]$ [37]. the probabilities of O1 and O2 are:

$$\omega_1 = \sum_{i=0}^t P_i \quad (2)$$

$$\omega_2 = \sum_{i=t+1}^{L-1} P_i = 1 - \omega_1 \quad (3)$$

The mean values of O1 and O2, respectively, are:

$$\mu_1 = \sum_{i=0}^t \frac{iP_i}{\omega_1} \quad (4)$$

$$\mu_2 = \sum_{i=t+1}^{L-1} \frac{iP_i}{\omega_2} \quad (5)$$

The mean grayscale value of the entire image is:

$$\mu = \sum_{i=0}^{t-1} iP_i = \omega_1\mu_1 + \omega_2\mu_2 \quad (6)$$

The variances of the two classes are:

$$\sigma_1^2 = \sum_{i=0}^t \frac{(i - \mu_1)^2 P_i}{\omega_1} \quad (7)$$

$$\sigma_2^2 = \sum_{i=t+1}^{L-1} \frac{(i - \mu_2)^2 P_i}{\omega_2} \quad (8)$$

The intra-class variance, inter-class variance and total variance of the two classes in OTSU are:

$$\sigma_W^2 = \omega_1\sigma_1^2 + \omega_2\sigma_2^2 \quad (9)$$

$$\sigma_B^2 = \omega_1(\mu_1 - \mu)^2 + \omega_2(\mu_2 - \mu)^2 = \omega_1\omega_2(\mu_2 - \mu_1)^2 \quad (10)$$

$$\sigma_T^2 = \sum_{i=0}^{L-1} (i - \mu_T)^2 P_i \quad (11)$$

The optimal threshold is determined when the variance between classes is maximum, and the formula is as follows:

$$T = \text{Max}\sigma_b^2(t) \quad (12)$$

OTSU algorithms is widely used in many fields, especially in pattern recognition and simulated vision. Some researchers have applied it to medical Magnetic Resonance Imaging, marine oil pollution detection and Computed Tomography image processing [38]. Sea ice recognition is actually to extract sea ice from remote sensing images, but the background in remote sensing images may be more complicated, and there may be no strict double peaks, and it is difficult to determine a reasonable threshold.

4.2 Recognition of Bohai Sea Ice based on Random Forests

Nowadays, when using satellite data information, the meteorological business usually selects only the data under clear sky conditions and eliminates all the fields of view or data polluted by clouds, this processing method is relatively simple, but the elimination of clouds will waste a large amount of satellite observation data. In practice, the proportion of the basic cloud-free situation in the Bohai Sea region is relatively small. In particular, MODIS data such as data acquisition frequency of 1–2 days is likely to result in the rejection of dozens of days in a year, and the waste of data appears more serious.

The multi-channel dynamic threshold method for sea ice identification, which cannot achieve sea ice identification under cloud coverage, and the Bohai Sea region also appears to be fully covered by clouds, and there is continuous cloud cover over Liaodong Bay in winter, and this situation cannot

achieve detection of sea ice. In fact, sea ice recognition is also the classification of sea ice and other interference objects. The classification method can be used to train and model the input data and classify the sea ice classification results.

The random forests algorithm uses decision trees as the base learner for integrated learning, where the decision tree is a tree-like data structure composed of root nodes, intermediate nodes and leaf nodes. The process of the algorithm is to first use the Bagging algorithm to randomly draw data samples from the sample training set D to obtain k independent training sample sets $\{D_1, D_2, \dots, D_k\}$ and then construct the corresponding decision trees according to the different training sample sets. k sample sets are generated k decision trees $\{F_1, F_2, \dots, F_k\}$. The criterion for selecting node features for the decision tree is the Gini coefficient, and if there are N category instances in the training set D , the Gini coefficient is calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^k [P(i) * P(i)] \quad (13)$$

where, $P(I)$ represents the proportion of class I sample data in the data set on the current node. When the training sample set D is divided into two subsets D_1 and D_2 by using the characteristic attribute F , the definition formula of Gini coefficient is:

$$Gini = Gini(D) - Gini(D_1) - Gini(D_2) \quad (14)$$

For each CART (classification and regression tree) decision tree, the training sample subset is started from the root node for training. If the training reaches the termination condition of splitting, the current node is set as a leaf node. If the termination condition of the split is not reached, then use the Gini coefficient to select an optimal feature from the N -dimensional features, and then divide the sample data on the current node into the left and right child nodes, and then continue to train other untrained nodes. Until all nodes are trained or are marked as leaf nodes. After training all CART, each tree can predict the test sample data set according to the node threshold, and use the voting method to determine the final classification result of the entire random forests based on the classification results of each tree. The features generated by using the decision tree alone usually have the problem of overfitting. Random forests first calculate the importance of features and then performs feature selection. The goal of random forests features selection is to select a small number of features that can meet the requirements of use, preventing overfitting. The sea ice in the Bohai Sea changes with the seasons, and some spectral channels of FY4A are also affected by solar radiation, so the use of random forests for feature selection is more suitable for practical application scenarios. There are many new algorithms proposed based on random forests such as XGboost (A machine learning method) [39]. The time complexity and space complexity of XGboost is much higher than that of random forests. XGboost needs to store not only the feature values but also the index values of the gradient statistics of the samples corresponding to the features. Therefore, random forests deployment on lightweight servers has better efficiency. In practical application scenarios, the Bohai Sea ice recognition algorithm usually runs on portable devices during field operations, and cannot run models with high complexity.

This article is based on FY-4A data and 6132 pixels in the Bohai Sea to classify the Bohai Sea ice. For FY-4A data and multi-channel adaptive thresholding method of Bohai Sea ice identification the binary data of Bohai Sea Ice identification results of the adjacent 24-hour multi-channel adaptive thresholding method are used as input and labeled as the sea ice operational data of the 24th hour.

5 Experiment and Result

5.1 Evaluation Method

The FY4A geostationary meteorological satellite started work in 2018 and began to provide stable quality product data in 2019. In this paper, we use all available data for three years from 2019 to 2021, and divide the data into 3 folds by year. To prevent the occurrence of over-fitting and under-fitting problems of the model and evaluate the generalization ability of the model, the model training in this paper adopts the 3-fold cross-validation method.

In the experiments of this chapter, multiple evaluation criteria are used to evaluate the random forests sea ice classification model. In this paper, a binary confusion matrix is used as shown in Tab. 2. The confusion matrix can clearly see the sample classification and is used to calculate the evaluation index of the classification.

Table 2: Classification result confusion matrix

		Predicted results	
		Positive	Negative
Real result	Positive	TP (true positive)	FN (false negative)
	Negative	FP (false positive)	TN (true negative)

Accuracy rate: The ratio of the number of samples with correct prediction results to the total number of samples, the calculation formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

Precision rate: Among all the samples predicted to be sea ice, the ratio of sea ice samples that are predicted to be correct samples is calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

Recall rate: The ratio of samples that are correctly predicted to be sea ice to all samples that are actually sea ice is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

F1-score: It is related to the precision rate and recall rate, which can better express the classification effect of the model. The calculation formula is:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (18)$$

5.2 Multichannel Adaptive Threshold Algorithm

The results of applying the OTSU dynamic thresholding algorithm for sea ice identification are shown in Fig. 3, and two days of experimental data on January 3, 2020 and February 12, 2020 are selected for demonstration, where yellow indicates land, blue indicates ocean, and white indicates sea ice. Figures (a) and (d) in Fig. 3 represent the sea ice identification results based on the Multichannel

adaptive threshold method, figures (b) and (e) represent the albedo-bright temperature threshold method identification results, and figures (c) and (f) represent the sea ice business. As can be seen from the figure, the Multichannel adaptive thresholding algorithm for sea ice identification in the Bohai Sea greatly improves the sea ice identification results compared with the albedo -bright temperature thresholding method. The Multichannel adaptive thresholding method is more suitable for Bohai Sea ice identification under clear sky conditions, and the dynamic threshold selection by dynamic threshold algorithm can reduce the workload of manual threshold adjustment and selection caused by seasonal change variation, which also shows that FY-4A data can be applied to Bohai Sea ice identification.

Our process for identifying sea ice in the Bohai Sea using adaptive thresholding is as follows:

- 1): Input the data from channel 2, 3, 5, 12 and 13, and calculate IST (Ice Surface Temperature) and NDSI (Normalized Difference Snow Index);
- 2): Normalize channel 2, IST and NDSI data to the 0–100 range
- 3): Interval alienation of channel 2, IST and NDSI data
- 4): Interval quantification of channel 2, IST and NDSI data
- 5): Set high and low thresholds for reflectance, IST, and NDSI, and use the thresholds for recognition, respectively. If the two recognition results are consistent, end; if not, proceed to the next step;
- 6): Determine dynamic thresholds. For the pixels with inconsistent recognition results, all the inter-class variances are calculated in sequence according to OTSU, and finally the inter-class variances are compared to determine a new threshold.

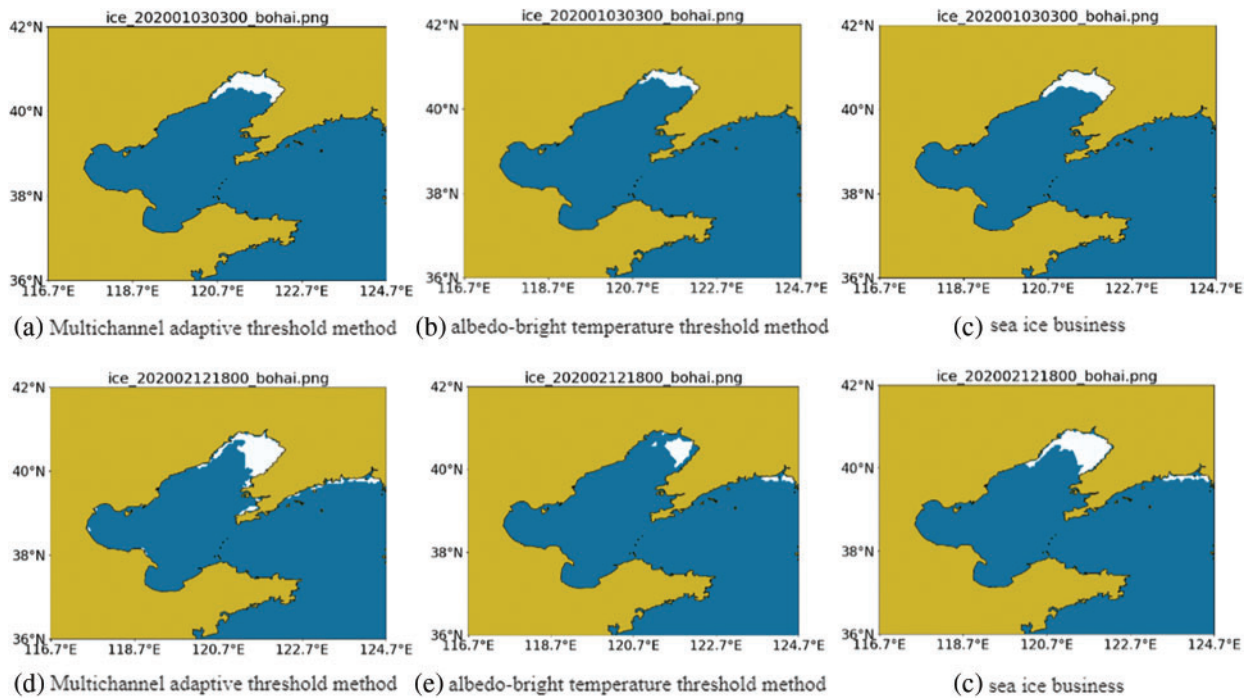


Figure 3: Comparison of recognition results

5.3 Random Forests

This paper is based on the random forests classification model to train and verify the Bohai Sea ice recognition model. The performance of the random forests classification model is mainly affected by the number of decision trees and the maximum depth of the trees, so this paper focuses on these two parameters for the study of tuning the parameters. A larger number of policy trees can lead to a better and more stable performance of the model, but it also makes the training slower. The tuning process uses a cross-validation method for training and validation, and the average score of cross-validation is used as the evaluation criterion. The learning curve is used to find the optimal value to correct the classification accuracy to a higher level. The number of decision trees and the maximum depth of the model inputs are parametrized separately, with the number of decision trees tested at 10 intervals from 1 to 400 and the maximum depth tested at 1 interval from 1 to 50 to determine the optimal parameter configuration for both input models. The workflow chart of the algorithm in this paper is shown in Fig. 4

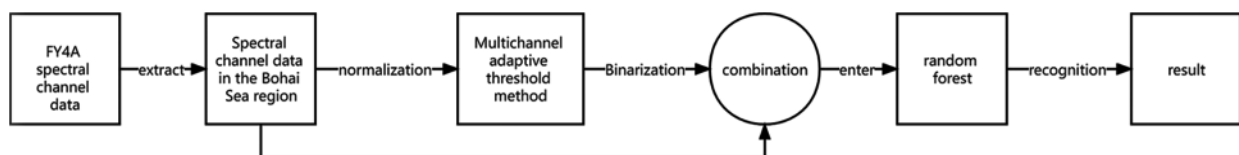


Figure 4: Workflow chart

Firstly, the model input is tuned to the sea ice identification results, and the learning curve of the specific tuning process is shown in the following Figs. 5 and 6. The number of decision trees was tested at intervals of 10 from 1 to 400, starting with an approximate range of 50 to 70, where the best number of decision trees was derived as 61 with a maximum score of 0.91992762. After further adjusting the parameters within this range, it is concluded that the number of optimal decision trees is 55, and the maximum score is 0.92009075. The model score has been improved. The maximum depth of the tree was tested at intervals from 1 to 50, and the optimal depth of the model was experimentally determined to be 5.

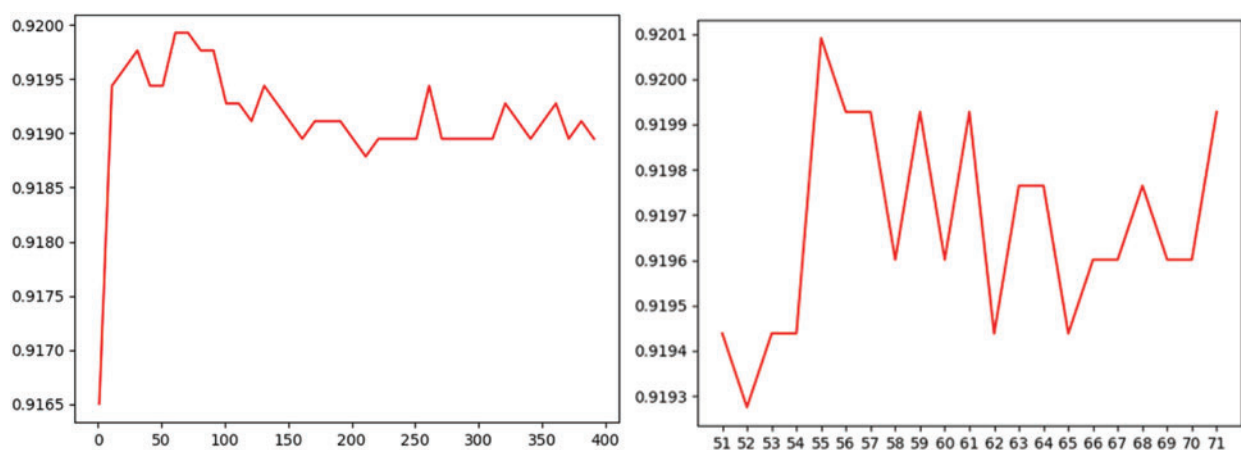


Figure 5: Quantity learning curve of decision tree

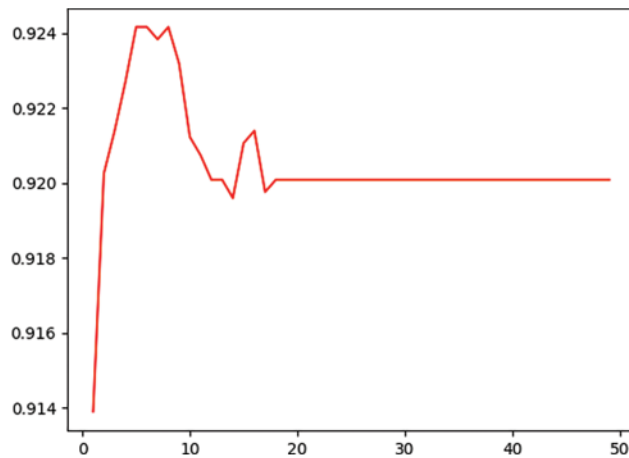


Figure 6: Decision tree maximum depth learning curve

5.4 Experimental Results

The experimental results were compared and analyzed with the results of sea ice operational data. Accuracy rate, Precision rate, Recall rate and F1-score were calculated by TP, FP, TN and FN. The sea ice in the Bohai Sea only exists in winter. We trained the model on the balanced and imbalanced datasets, respectively. The results show that the problem of imbalanced data can be overcome by using the random forests method. The experimental results are selected from the experimental data on December 25, 2020, as shown in the following [Tabs. 3, 4](#) and [Fig. 7](#)

Table 3: TP; FP; TN; FN

Time	Balanced/ imbalanced	Model	TP	FP	FN	TN
2020-12-25 00:00:00	imbalanced	Multichannel adaptive threshold method (MATM)	718	109	196	5109
		Random Forests (RF)	538	68	376	5150
	balanced	Multichannel adaptive threshold method (MATM)	829	379	85	4839
		Random Forests (RF)	590	125	324	5093
2020-12-25 06:00:00	imbalanced	Multichannel adaptive threshold method (MATM)	154	230	760	4988
		Random Forests (RF)	465	59	449	5159
	balanced	Multichannel adaptive threshold method (MATM)	370	321	544	4897
		Random Forests (RF)	535	111	379	5107

(Continued)

Table 3: Continued

Time	Balanced/ imbalanced	Model	TP	FP	FN	TN
2020-12-25 12:00:00	imbalanced	Multichannel adaptive threshold method (MATM)	235	41	679	5177
		Random Forests (RF)	462	65	452	5153
	balanced	Multichannel adaptive threshold method (MATM)	623	332	291	4886
		Random Forests (RF)	596	115	318	5103
2020-12-25 18:00:00	imbalanced	Multichannel adaptive threshold method (MATM)	210	21	704	5197
		Random Forests (RF)	470	47	444	5171
	balanced	Multichannel adaptive threshold method (MATM)	459	197	455	5021
		Random Forests (RF)	564	103	350	5115
2020-12-25 24:00:00	imbalanced	Multichannel adaptive threshold method (MATM)	594	39	320	5179
		Random Forests (RF)	528	50	386	5168
	balanced	Multichannel adaptive threshold method (MATM)	775	208	139	5010
		Random Forests (RF)	580	112	334	5106

Table 4: Accuracy rate; Precision rate; Recall rate; F1-score

Time	Balanced/ imbalanced	Model	Accuracy rate	Precision rate	Recall rate	F1-score
2020-12-25 00:00:00	imbalanced	Multichannel adaptive threshold method (MATM)	95.03	86.82	78.56	82.48
		Random Forests (RF)	92.76	88.78	58.86	70.79
	balanced	Multichannel adaptive threshold method (MATM)	68.63	90.70	92.43	78.13

(Continued)

Table 4: Continued

Time	Balanced/ imbalanced	Model	Accuracy rate	Precision rate	Recall rate	F1-score
2020-12-25 06:00:00	imbalanced	Random Forests (RF)	82.52	64.55	92.68	72.44
		Multichannel adaptive threshold method (MATM)	83.86	40.10	16.85	23.73
	balanced	Random Forest (RF)	91.70	88.14	51.20	64.78
		Multichannel adaptive threshold method (MATM)	53.55	52.30	84.95	50.88
2020-12-25 12:00:00	imbalanced	Random Forests (RF)	84.05	58.53	92.01	68.59
		Multichannel adaptive threshold method (MATM)	88.26	85.14	25.71	39.50
	balanced	Random Forests (RF)	91.36	86.23	50.00	63.30
		Multichannel adaptive threshold method (MATM)	63.67	65.97	89.32	64.80
2020-12-25 18:00:00	imbalanced	Random Forests (RF)	83.8256	65.21	92.94	73.35
		Multichannel adaptive threshold method (MATM)	88.18	90.91	22.98	36.68
	balanced	Random Forests (RF)	91.39	91.06	46.83	61.85
		Multichannel adaptive threshold method (MATM)	68.27	52.95	89.32	59.64
2020-12-25 24:00:00	imbalanced	Random Forests (RF)	84.56	61.71	92.61	71.35
		Multichannel adaptive threshold method (MATM)	94.15	93.84	64.99	76.80
	balanced	Random Forests (RF)	92.89	91.35	57.77	70.78
		Multichannel adaptive threshold method (MATM)	78.84	84.79	94.34	81.71
		Random Forests (RF)	83.82	63.46	92.73	72.23

Comparing the two models, we can find that the random forests model has better classification results and smoother performance. The result plots are shown below, and (a), (b), (c) and (d) in Fig. 8 show the Multichannel adaptive threshold method results, random forests identification results, MODIS sea ice product results and sea ice business results, respectively. From the figure, it can be seen that: under non-clear sky conditions, the sea ice identification results using the Multichannel adaptive threshold method are very poor due to cloud coverage, especially in the Liaodong Bay waters,

Bohai Bay and Laizhou Bay region. Compared with the results of the Multichannel adaptive threshold method and MODIS sea ice products, the random forests classification algorithm has improved, and is closer to the sea ice business results.

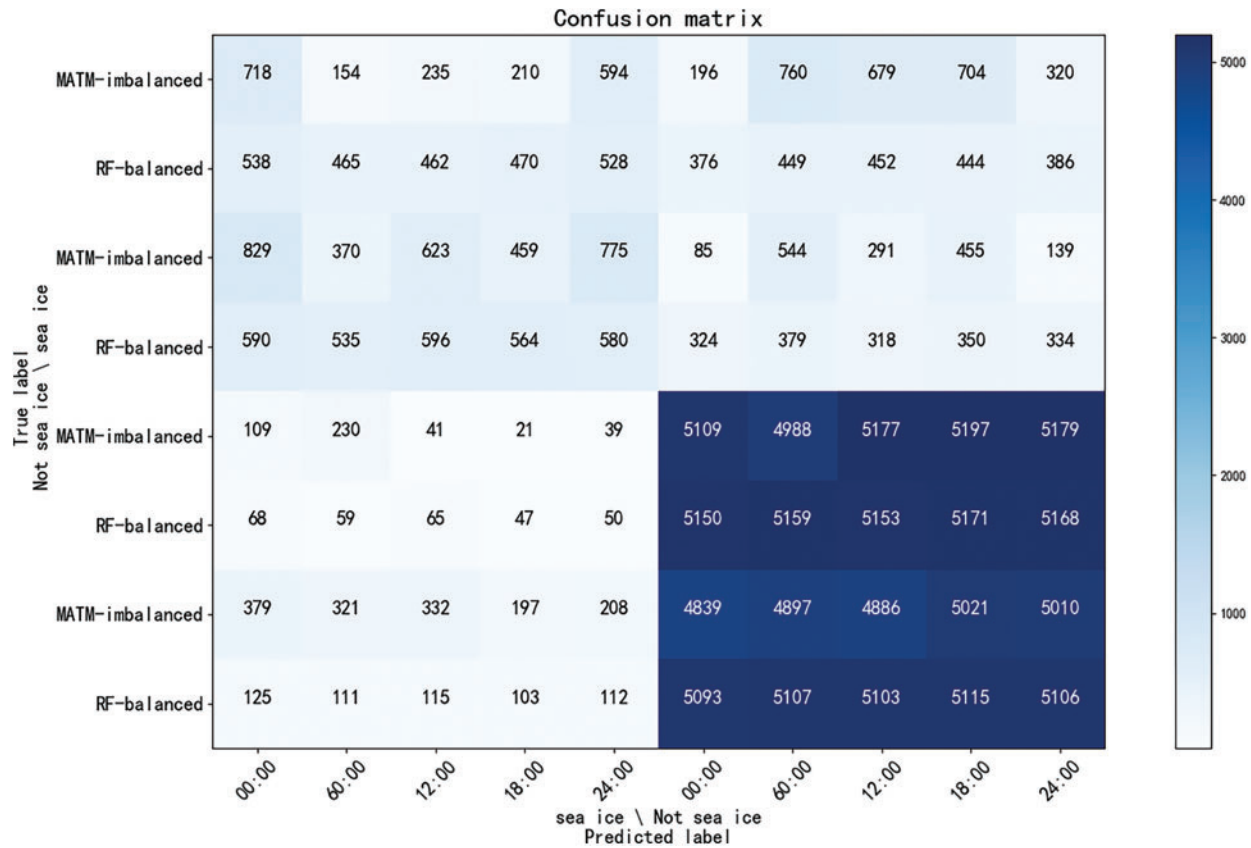


Figure 7: Confusion matrix

The random forests algorithm is close to the results of the sea ice business, and the recognition effect is better than that of the Multichannel adaptive threshold method; Compared with MODIS sea ice products, random forests greatly improve the effect of sea ice recognition. Therefore, through the analysis of the experimental results, the random forests algorithm is more suitable for the recognition of sea ice in the Bohai Sea. Feature extraction by random forests algorithm can reduce the workload of threshold adjustment caused by seasonal changes. It also shows that FY-4A data can be applied to sea ice recognition in the Bohai Sea. Due to the lack of satellite remote sensing data, when there is a large area of cloud coverage above the Bohai Sea, the effect of sea ice recognition by the Multichannel adaptive threshold method decreases significantly, and even the sea ice cannot be recognized. Because of this situation, using random forests can reduce the influence of clouds on sea ice identification.

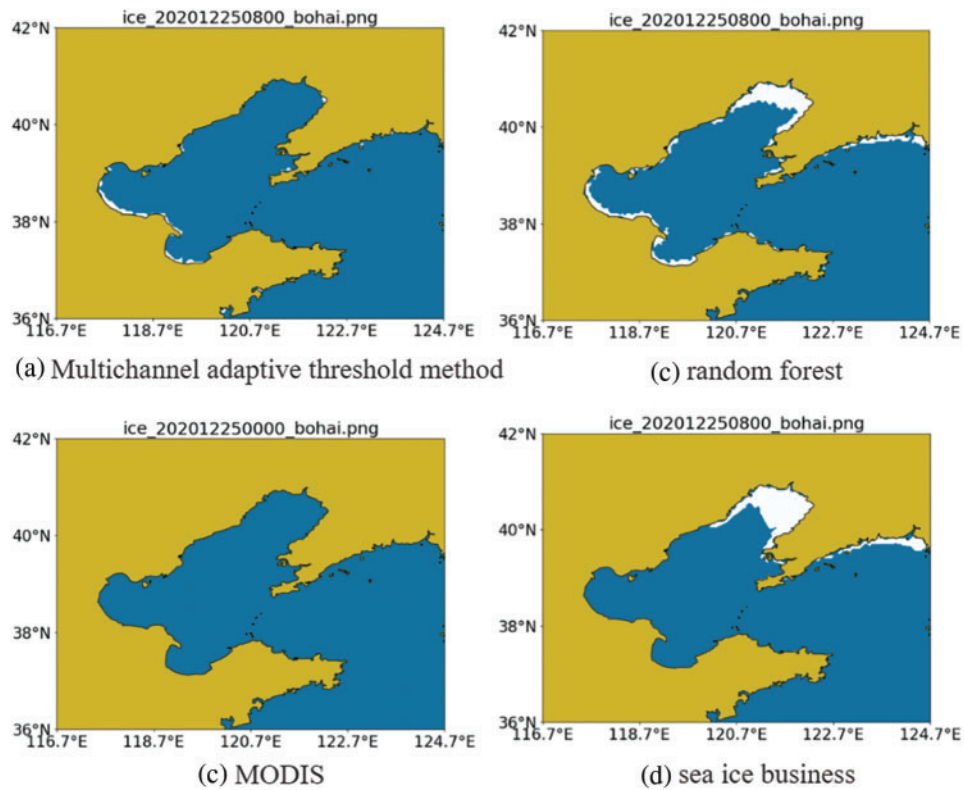


Figure 8: Comparison of recognition results

6 Conclusion

In this paper, the existing sea ice identification algorithms are studied and found to be basically for sea ice identification under clear sky conditions without considering the identification under cloud cover. Therefore, this paper combines the results of sea ice identification by the threshold method and designs and implements a random forests based sea ice classification algorithm for the Bohai Sea to achieve more effective sea ice identification results under clouds. The detection and removal of clouds in sea ice identification algorithm is crucial to sea ice identification and determines the accuracy of identification results. After the experiment, it can be seen that the data of FY-4A cloud detection product has some errors in the detection results of clouds, and the subsequent work can further research on the detection of clouds and improve the effect of sea ice identification. Geostationary weather satellites work on the same principle and have similar observation channels, so theoretically the random forests method can be used for most geostationary weather satellites. There are many traditional meteorological methods to achieve sea ice identification, and in the future, we intend to try new methods such as deep learning methods to identify sea ice to improve sea ice identification.

Funding Statement: This research was supported by the National Natural Science Foundation of China under Grant No.61772280 and No.62072249

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] Q. Tang, X. Zhang, X. Yang and J. A. Francis, “Cold winter extremes in northern continents linked to Arctic sea ice loss,” *Environmental Research Letters*, vol. 8, no. 1, pp. 14036, 2013.
- [2] A. K. Rennermalm, L. C. Smith, J. C. Stroeve and V. W. Chu, “Does sea ice influence greenland ice sheet surface-melt?,” *Environmental Research Letters*, vol. 4, no. 2, pp. 024011, 2009.
- [3] M. M. Holland and C. M. Bitz, “Polar amplification of climate change in coupled models,” *Climate Dynamics*, vol. 21, no. 3, pp. 221–232, 2003.
- [4] H. Su and Y. Wang, “Using MODIS data to estimate sea ice thickness in the Bohai Sea (China) in the 2009–2010 winter,” *Journal of Geophysical Research: Oceans*, vol. 117, no. 10, pp. 1–8, 2012.
- [5] A. D. Fraser, R. A. Massom and K. J. Michael, “A method for compositing polar MODIS satellite images to remove cloud cover for landfast sea-ice detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 9, pp. 3272–3282, 2009.
- [6] G. Riggs and D. A. Hall, “Reduction of cloud obscuration in the MODIS snow data product,” in *59th Eastern Snow Conf.*, Stowe, VT, USA, pp. 205–212, 2003.
- [7] G. McWilliams, M. Goldberg, L. Zhou, B. Reed and J. Weinrich, “Satellites providing critical support for environmental disaster monitoring and response: from TIROS-1, the united states’ first weather satellite, to the current generation of environmental satellites posters,” in *AGU Fall Meeting 2019. AGU*, San Francisco, CA, USA, 2019.
- [8] T. Heacock, T. Hirose, F. M. Manore and B. Ramsay, “Sea-ice tracking on the east coast of Canada using NOAA/AVHRR imagery,” *Annals of Glaciology*, vol. 17, pp. 405–413, 1993.
- [9] J. Key and M. Haefliger, “Arctic ice surface temperature retrieval from AVHRR thermal channels,” *Journal of Geophysical Research: Atmospheres*, vol. 97, no. 5, pp. 5885–5893, 1992.
- [10] R. W. Lindsay and D. A. Rothrock, “Arctic sea ice leads from advanced very high resolution radiometer images,” *Journal of Geophysical Research: Oceans*, vol. 32, no. 22, pp. 4533–4544, 2005.
- [11] J. D. Pinhirne, S. P. Palm and W. D. Hart, “Antarctica cloud cover for october 2003 from GLAS satellite lidar profiling,” *Geophysical Research Letters*, vol. 32, no. 22, pp. 1–4, 2005.
- [12] J. J. Yackel, V. Nandan, M. Mahmud, R. Scharien and J. W. Kang, “A spectral mixture analysis approach to quantify arctic first-year sea ice melt pond fraction using quickbird and MODIS reflectance data,” *Remote Sensing of Environment*, vol. 204, no. 10, pp. 704–716, 2018.
- [13] M. Mäkynen and J. Karvonen, “MODIS sea ice thickness and openwater-sea ice charts over the barents and kara seas for development and validation of sea ice products from microwave sensor data,” *Remote Sensing*, vol. 9, no. 12, pp. 1–38, 2017.
- [14] S. U. Islam, S. Jan, A. Waheed, G. Mehmood, M. Zareei *et al.*, “Land-Cover classification and its impact on peshawar’s land surface temperature using remote sensing,” *Computers, Materials & Continua*, vol. 70, no. 2, pp. 4123–4145, 2022.
- [15] H. Sallila, S. Farrell, J. McCurry and E. Rinne, “Assessment of contemporary satellite sea ice thickness products for Arctic sea ice,” *The Cryosphere*, vol. 13, no. 4, pp. 1187–1213, 2019.
- [16] S. Yuan, C. Liu, X. Liu, Y. Chen and Y. Zhang, “Research advances in remote sensing monitoring of sea ice in the bohai sea,” *Earth Science Informatics*, vol. 14, no. 4, pp. 1729–1743, 2021.
- [17] A. Asokan, J. Anitha, B. Patrut, D. Danculescu and D. J. Hemanth, “Deep feature extraction and feature fusion for bi-temporal satellite image classification,” *Computers, Materials & Continua*, vol. 66, no. 1, pp. 373–388, 2021.
- [18] J. Esteban, R. McRoberts, A. Fernández-Landa, J. Tomé and E. Nsset, “Estimating forest volume and biomass and their changes using random forests and remotely sensed data,” *Remote Sensing*, vol. 11, no. 16, pp. 1944, 2019.
- [19] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 1249, 2018.
- [20] Q. Wang, Y. Luo, H. Guo, P. Guo and J. Wei, “A pagerank-based wechat user impact assessment algorithm,” *Journal of New Media*, vol. 3, no. 2, pp. 53–62, 2021.

- [21] A. Assiri, "Anomaly classification using genetic algorithm-based random forest model for network attack detection," *Computers, Materials & Continua*, vol. 66, no. 1, pp. 767–778, 2021.
- [22] Z. Yu, C. Zhang, N. Xiong and F. Chen, "A new random forest applied to heavy metal risk assessment," *CSSE-Computer Systems Science and Engineering*, vol. 40, no. 1, pp. 207–221, 2022.
- [23] S. Wang, Z. Zhang, S. Geng and C. Pang, "Research on optimization of random forest algorithm based on spark," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3721–3731, 2022.
- [24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [25] M. A. Khan, S. A. Memon, F. Farooq, M. F. Javed, F. Aslam *et al.*, "Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest," *Advances in Civil Engineering*, vol. 2021, pp. 1–17, 2021.
- [26] V. Jackins, S. Vimal, M. Kaliappan and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021.
- [27] A. Y. Dawod and M. A. Sharafuddin, "Assessing mangrove deforestation using pixel-based image: A machine learning approach," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3178–3190, 2021.
- [28] N. Xu, S. Yuan, X. Liu, Y. Ma and W. Shi, "Risk assessment of sea ice disasters on fixed jacket platforms in Liaodong Bay," *Natural Hazards and Earth System Sciences*, vol. 20, no. 4, pp. 1107–1121, 2020.
- [29] X. He, N. Xu, X. Feng, X. Hu and H. Xu, "Assessing radiometric calibration of FY-4A/AGRI thermal infrared channels using cris and iasi," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
- [30] W. E. Dietrich, C. J. Wilson, D. R. Montgomery and J. McKean, "Analysis of erosion thresholds, channel networks and landscape morphology using a digital terrain model," *Journal of Geology*, vol. 101, no. 2, pp. 259–278, 1993.
- [31] C. Li, J. Ma, P. Yang and Z. Li, "Detection of cloud cover using dynamic thresholds and radiative transfer models from the polarization satellite image," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 222, pp. 196–214, 2019.
- [32] D. Grosvenor, O. Sourdeval, P. Zuidema, A. Ackerman, M. Alexandrov *et al.*, "Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives," *Reviews of Geophysics*, vol. 56, no. 2, pp. 409–453, 2018.
- [33] F. Zhou, "Inversion and accuracy verification of spectral calibration parameters for spaceborne infrared hyperspectral instrument," China: M.S. dissertation, Chinese Academy of Meteorological Sciences, 2019.
- [34] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 92, no. 6, pp. 1–16, 2021.
- [35] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems Man & Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [36] H. Du, X. Chen and J. Xi, "An improved background segmentation algorithm for fringe projection profilometry based on Otsu method," *Optics Communications*, vol. 453, no. 124206, pp. 1944, 2019.
- [37] Y. Quan, J. Sun, Y. Zhang and H. Zhang, "The method of the road surface crack detection by the improved Otsu threshold," in *2019 IEEE Int. Conf. on Mechatronics and Automation (ICMA)*, Tian Jing, TJ, China, pp. 1615–1620, 2019.
- [38] L. Zhang, L. Wang, Y. Huang and H. Chen, "Segmentation of thoracic organs at risk in CT images combining coarse and fine network," *In SegTHOR@ISBI*, vol. 11, no. 16, pp. 2–4, 2019.
- [39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. of the 22nd Acm Sigkdd Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, pp. 785–794, 2016.