agriculture

MDPI

Article

Mapping Soybean Planting Areas in Regions with Complex Planting Structures Using Machine Learning Models and Chinese GF-6 WFV Data

Bao She [1,2], Jiating Hu [1], Linsheng Huang [2,*], Mengqi Zhu [1] and Qishuo Yin [2]

[1] School of Spatial Informatics and Geomatics Engineering, Anhui University of Science and Technology, Huainan 232001, China; shebao518@aust.edu.cn (B.S.); 2021201668@aust.edu.cn (J.H.); 2020201379@aust.edu.cn (M.Z.)
[2] National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei 230601, China; p21301210@stu.ahu.edu.cn
* Correspondence: 06035@ahu.edu.cn; Tel.: +86-133-5569-0701

Abstract: To grasp the spatial distribution of soybean planting areas in time is the prerequisite for the work of growth monitoring, crop damage assessment and yield estimation. The research on remote sensing identification of soybean conducted in China mainly focuses on the major producing areas in Northeast China, while paying little attention to the Huang-Huai-Hai region and the Yangtze River Basin, where the complex planting structures and fragmented farmland landscape bring great challenges to soybean mapping in these areas. This study used Chinese GF-6 WFV imagery acquired during the pod-setting stage of soybean in the 2019 growing season, and two counties i.e., Guoyang situated in the northern plain of Anhui Province and Mingguang located in the Jianghuai hilly regionwere selected as the study areas. Three machine learning algorithms were employed to establish soybean identification models, and the distribution of soybean planting areas in the two study areas was separately extracted. This study adopted a stepwise hierarchical extraction strategy. First, a set of filtering rules was established to eliminate non-cropland objects, so the targets of subsequent work could thereby focus on field vegetation. The focal task of this study involved the selection of well-behaved features and classifier. In addition to the 8 spectral bands, a variety of texture features, color space components, and vegetation indices were employed, and the ReliefF algorithm was applied to evaluate the importance of each candidate feature. Then, a SFS (Sequential Forward Selection) method was applied to conduct feature selection, which was performed coupled with three candidate classifiers, i.e., SVM, RF and BPNN to screen out the features conductive to soybean mapping. The accuracy evaluation results showed that, the soybean identification model generated from SVM algorithm and corresponding feature subset outperformed RF and BPNN in both two study areas. The Kappa coefficients of the ground samples in Guoyang ranged from 0.69 to 0.80, while those in Mingguang fell within the range of 0.71 to 0.76. The near-infrared band (B4) and red edge bands (B5 and B6), the 'Mean' texture feature and the vegetation indices, i.e., EVI, SAVI and CI_{green}, demonstrated advantages in soybean identification. The feature selection operation achieved a balance between extraction accuracy and data volume, and the accuracy level could also meet practical requirements, showing a good application prospect. This method and findings of this study may serve as a reference for research on soybean identification in areas with similar planting structures, and the detailed soybean map can provide an objective and reliable basis for local agricultural departments to carry out agricultural production management and policy formulation.

Keywords: soybean; crop mapping; GF-6 WFV; machine learning; feature selection

check for updates

Citation: She, B.; Hu, J.; Huang, L.; Zhu, M.; Yin, Q. Mapping Soybean Planting Areas in Regions with Complex Planting Structures Using Machine Learning Models and Chinese GF-6 WFV Data. *Agriculture* **2024**, *14*, 231. https://doi.org/10.3390/agriculture14020231

Academic Editors: Dionissios Kalivas, Thomas Alexandridis, Konstantinos X. Soulis and Emmanouil Psomiadis

Received: 27 November 2023
Revised: 23 January 2024
Accepted: 25 January 2024
Published: 31 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soybean (*Glycine max* L.) is rich in oil and protein, making it a major source of edible oil and an important raw material for food industry and animal feed. It holds significant

value for industries such as oil production, food processing, and animal husbandry. China, as the homeland of soybeans, is also one of the world's most important soybean-producing countries. In 2021, the planting area reached 8415.4 kha, with a production of 16,395 kt, ranking among the top five in the world (from the "2022 China Rural Statistical Yearbook"). China is divided into five major soybean-producing regions according to climate conditions, soybean variety characteristics, and cultivation systems, among which the Northeast (spring-sown soybeans) and the Huang-Huai-Hai region (summer-sown soybeans) are the most important in terms of cultivated area and yield. In recent years, rapid economic development and improved living standards have led to a continuous growth in demand for soybeans. However, the domestic supply of soybeans is severely inadequate, resulting in a long-term dependence on imports of over 80%. Soybean has become the crop with the lowest self-sufficiency rate in China. Given the challenging soybean supply situation in the country, there is an urgent need to grasp detailed, accurate, and objective information about domestic soybean production and formulate policies in time to expand cultivation.

Timely comprehension of the distribution of crop planting areas is the basis for tasks such as crop growth monitoring, damage assessment and yield estimation. Traditionally, a combination of sampling statistics and manual surveys has been employed to estimate the cultivated area of specific crop. However, this approach is not only time-consuming and inefficient, but also susceptible to subjective factors in practice, and the statistical results cannot provide its spatial distribution. Remote sensing is an effective and widely applied technical means of synchronously capturing crop cultivation information over a large area. The monitoring results are not only timely and cost-effective, but also provide sufficient spatial details and meet the requirements of objectivity. Large-scale soybean mapping or planting areas extraction can be achieved through two strategies: hard classification and soft classification. The former primarily focuses on mapping the spatial distribution to obtain complete coverage of soybean fields and estimating its cultivated area. The latter delves into the interior of pixels and directly captures soybean abundance or percentage content at the pixel level, this method is advantageous for effectively addressing "mixed pixel" issue and provides more accurate estimation of soybean acreage [1,2].

Soybean mapping at the pixel level is typically based on single-date, multi-temporal, or time-series image data, by means of classification approach. The classification methods that can be employed include traditional supervised and unsupervised classifiers, e.g., maximum likelihood classification [3–5], primarily relying on the specific spectral characteristics of soybeans. The decision tree classification method is based on knowledge and involves establishing a set of filtering rules, the core of the algorithm is to select unique signatures of soybean, distinct from other land cover or crop types, and then construct corresponding rules for its definition. Decision tree is often built based on vegetation indices or band reflectance [6–8]. Some scholars have also developed specific remote sensing indices and combined them with other discriminant conditions to achieve soybean identification. Huang et al. [9] for instance, designed three phenology-based remote sensing indices based on the variations of soybean canopy moisture and chlorophyll content throughout its growing season and applied them to soybean mapping in Northeast China and four states in the United States, achieving good results. Chen et al. [10] analyzed time-series Sentinel-2 data and found that soybean exhibited both high NDVI (Normalized Difference Vegetation Index) (indicating high greenness) and high SWIR (Short-Wave Infrared) reflectance (representing relatively lower canopy moisture) during the peak growing stage, they developed the GWCCI (Greenness and Water Content Composite Index) index accordingly, which could effectively enhance the signal of soybean fields and achieve rapid mapping of large-scale soybean planting areas, through a simple threshold segmentation adopting single-date imagery. Object-oriented classification has been increasingly applied to crop distribution mapping, it first segments the image to create small patches (objects) with uncertain cover type attribution, and then defines their attribution based on specified spectral, textural, and morphological features. In practical implementation, it is often combined with other classification methods to extract soybean distribution, e.g., object-oriented classi-

fication with decision tree [11], or with multi-layer neural network [12]. The time-series analysis method draws inspiration from hyperspectral data processing, using time series of vegetation indices e.g., NDVI or EVI (Enhanced Vegetation Index) as input and employing waveform matching algorithms like Spectral Angle Mapper (SAM) classifier, and soybean planting areas could also be extracted [13,14].

Machine learning methods are widely applied in crop mapping due to their practicality. This approach involves establishing a feature space based on various forms of feature datasets, and automatically build identification model for specified cover types with the support of training samples, enabling automatic feature selection. Common machine learning algorithms include Artificial Neural Networks (ANN), Random Forest (RF), and Support Vector Machine (SVM) etc. Soybean identification based on machine learning methods can utilize various input data forms including spectral bands [15,16], multiple vegetation indices [17], and SAR polarization features [18]. Liang et al. [19] found that the red edge bands and the red edge vegetation indices generated from Chinese GF-6 WFV data had significant advantages in soybean mapping. Ren et al. [20] combined Sentinel-1/2 SAR (Synthetic Aperture Radar) and multispectral images and used the RF algorithm to classify major crops in the mountainous areas of Inner Mongolia, finding that the red edge and shortwave infrared bands exhibited superiority in maize and soybean classification. Some studies introduce texture features to the extraction of soybean fields. Shen et al. [21] used RF algorithm to extract the distribution of maize and soybean planting areas in Kedong County, Heilongjiang Province, and Sixian County, Anhui Province, based on GF-1 WFV data acquired in the mid-to-late growth stage of soybean and maize (mid to late August), confirming the effectiveness of Second Moment and Entropy in soybean mapping in the two study areas. In addition, multiple key phenological metrics closely related to the soybean growth cycle can be derived from the temporal profile of vegetation indices, and taken as input for machine learning algorithms to achieve soybean mapping [22]. Moreover, machine learning schemes can be enhanced by implementing certain measures to improve their performance. For instance, conducting feature selection for different machine learning algorithms to identify discriminative features for soybean identification has been proven effective in obtaining better extraction results [23,24]. Additionally, employing ensemble learning strategies based on multiple machine learning algorithms can leverage the complementarity between different algorithms, thereby refining the initial results [25]. As a subset of machine learning methods, deep learning networks, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), possess strong information mining capabilities and can achieve higher extraction accuracy compared to traditional machine learning classifiers [26,27].

As for the issue related to soybean identification and planting areas extraction, although many scholars at home and abroad have explored and achieved a batch of results, existing research still mainly focuses on countries or regions with large-scale and highly mechanized soybean cultivation and relatively simple planting structures (e.g., the major soybean producing areas in Northeast China). However, in regions with diverse crop types, complex planting structures, and fragmented farmland landscapes, e.g., the Huang-Huai-Hai region and the Yangtze River Basin in China, the planting conditions or natural environments covered in previous studies may not be typical or representative enough. Therefore, further investigation is needed to assess the applicability and generalizability of the soybean identification models constructed for the aforementioned regions. Furthermore, soybean cultivated in Southern China as well as the provinces spanning the geographical North and South China accounts for over 30% of the national total planting scale (e.g., 32% in 2021 according to the "2022 China Rural Statistical Yearbook"). However, there have been few studies conducted on soybean mapping in these areas, particularly with the application of domestic satellite images, such as the Chinese GaoFen series. With respect to research methods, machine learning has become the mainstream approach owing to its high efficiency, ideal extraction results, and easy operability. Nevertheless, there is a lack of comparative analysis regarding the application effects of machine learning method in

soybean identification in the Huang-Huai-Hai region and the Yangtze River Basin. Given the deficiencies in existing research, this study selected typical soybean-producing counties in the northern plain of Anhui Province (belonging to the Huang-Huai-Hai region) and the Jianghuai hilly region (part of the Yangtze River Basin) as the study areas, and adopted Chinese GF-6 WFV image acquired during the pod-setting stage to carry out soybean mapping. The main objectives of this study include: (1) investigating the application effects of machine learning algorithms in soybean mapping in the two regions and establishing reliable models for soybean identification; (2) performing the selection of optimum features and classifier for each study area to achieve model optimization. The findings of this paper may provide a reference for related studies on fine-grained soybean identification under similar planting conditions.

2. Materials and Methods

2.1. Study Area

Anhui Province is situated in the eastern part of China, spanning a geographical scope of 29°41′~34°38′ N and 114°54′~119°37′ E. It covers a land area of 140,100 km^2 and is characterized by the presence of two major river systems, i.e., the Yangtze River and the Huaihe River. The province exhibits several distinct landforms, comprising the northern plain, the hills situated between the Yangtze River and the Huaihe River, the plain along the Yangtze River, and the western and southern mountainous areas, with significant differences in natural conditions. Anhui spans the north and south China in geographical sense, the region north of the Huaihe River is characterized by a warm temperate semi-humid monsoon climate, while the region south of the Huaihe River has a subtropical humid monsoon climate. Anhui Province is an important national production base for grain and oil crops, with soybean cultivation consistently ranking among the top three in the country. Soybean is mainly grown in the northern plain, which has a low-lying terrain and mainly cultivates dryland crops. The Jianghuai hilly region, characterized by hilly and rough terrain with a highly fragmented land surface, is dominated by rice cultivation. In this study, two county-level administrative units, i.e., Guoyang in the northern plain and Mingguang in the Jianghuai hilly region were selected as the study areas. These regions have diverse crop types, with the common practice of mixed cropping of different crops, leading to fragmented farmland landscapes.

Guoyang is located in the northern plain of Anhui Province, spanning from 33°27′ to 33°47′ N and 115°53′ to 116°33′ E, with a total area of 2110 km^2. The terrain here is gentle, slightly higher in the northwest and lower in the southeast, with an average elevation ranging from 26.5 to 33.5 m. Guoyang has a mild climate with similar distribution of rainfall and heat throughout the year. The annual mean temperature is 15.1 °C, annual precipitation is 851.6 mm, and there are approximately 2015.7 h of sunshine per year. The county has a total of 155.4 kha of arable land, and its major summer crops include soybean, corn, sorghum, sweet potato, peanut, as well as vegetables and medicinal herbs. The soybean planting scale in Guoyang has consistently ranked first in the province., with an area of 68,029 ha in 2021 (according to the "2022 Statistical Yearbook of Bozhou City"). The planting pattern is drastically different between the north and south of the Guohe River. To the north of this river, soybean and corn are predominantly mixed planted with soybean holding a significantly higher proportion than corn. In contrast, to the south of the river, corn dominates the agricultural landscape (Figure 1).

Mingguang is located in the Jianghuai hilly region, covering the spatial range between 32°27′~33°13′ N and 117°56′~118°25′ E, with an area of 2350 km^2. The terrain in Mingguang varies from south to north, encompassing low mountains in the south, hilly areas in the central region, and plains in the north. The average elevation is 44.8 m, and the Nushan Lake lies in the northern part of this county. The annual mean temperature is 15 °C, annual rainfall amounts to 934 mm, and there are 2200 h of sunshine per year, the region experiences hot summer with abundant rainfall occurring during the same period. Mingguang has 110.6 kha of arable land, with major summer crops including rice, soybean,

corn, peanut, sweet potato, vegetables and medicinal herbs. Mingguang always holds the top position among all county-level administrative units in the Jianghuai region in terms of soybean acreage.

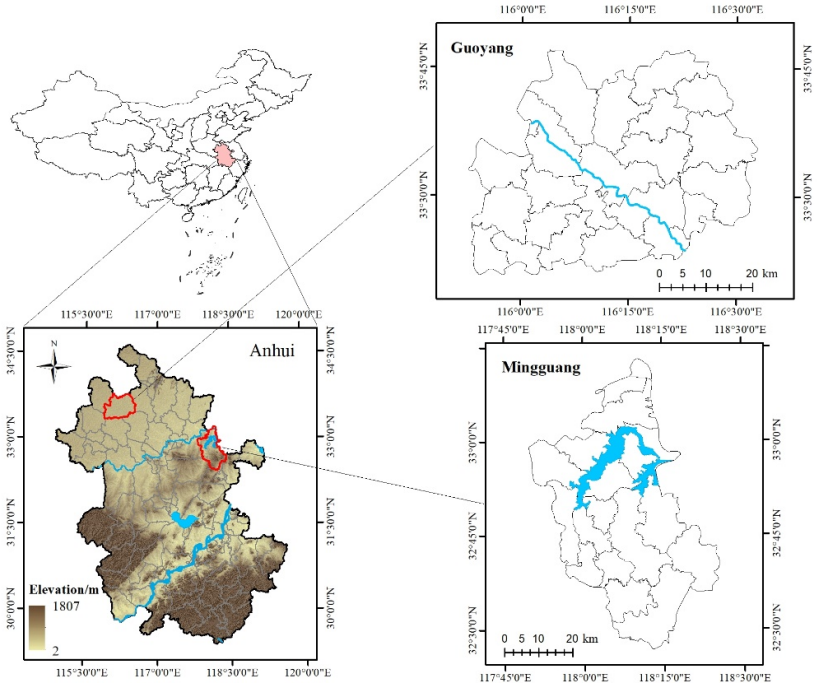


Figure 1. Map of the study area.

The soybean in Guoyang is usually sown in mid to late June, begins podding in mid-August, and undergoes harvesting from late September to early October (China Meteorological Data Service Centre, http://data.cma.cn/, accessed on 22 February 2019), the major phenological stages are illustrated in Figure 2. It is noteworthy that Mingguang lacks a dedicated crop growth and development observation station, and the data observed at the nearby Chuzhou station does not include information specific to soybean crop. After consulting with the Mingguang Agricultural Technology Extension Center, we learn that soybean in the local area is generally sown in mid-June, enters pod-setting stage in early August, and is harvested in late September, the growing season is relatively shorter compared to Guoyang.

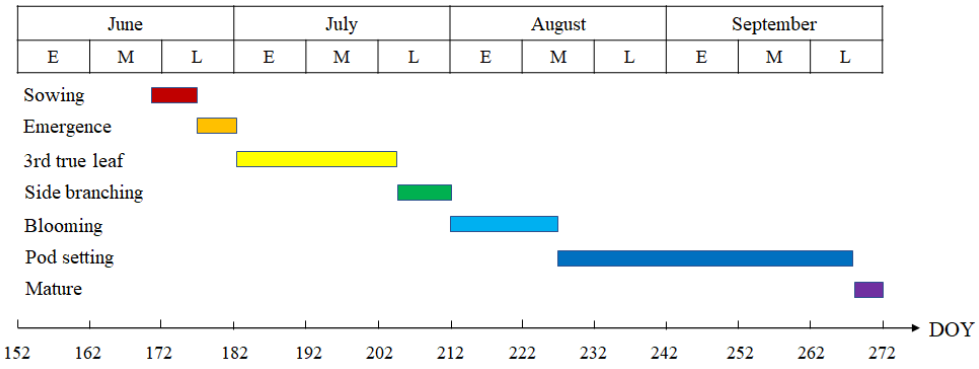


Figure 2. Time span of the major phenological phases of soybean in Guoyang.

2.2. Data

This study employed GF-6 WFV image as the remote sensing data source. The Chinese GF-6 satellite was launched in June 2018 and became operational in March of the following

year. It is equipped with a 2-m panchromatic/8-m multispectral (PMS) high-resolution camera and a 16-m multispectral medium-resolution wide field-of-view (WFV) camera. The WFV has eight working bands (Table 1) with an observation width of 800 km and offers a 4-day revisit cycle. It is the first domestic satellite that provides red-edge bands (710 nm, 750 nm), and is well-suited for precision agricultural monitoring, hence it is referred to as the "Agricultural Satellite". This study employed its L1A product, sourced from the China Centre for Resources Satellite Data and Application (https://data.cresda.cn/#/home). The image was acquired on 16 August 2019, during the pod-setting stage of soybeans, a time that has been shown in numerous studies to be suitable for soybean mapping using remote sensing datasets [18,28,29].

Table 1. Band configuration of the GF-6 WFV instrument.

Band Number	Band Name	Spectral Range/μm
B1	Blue	0.45~0.52
B2	Green	0.52~0.59
B3	Red	0.63~0.69
B4	Near infrared/NIR	0.77~0.89
B5	Red edge-1	0.69~0.73
B6	Red edge-2	0.73~0.77
B7	Purple	0.40~0.45
B8	Yellow	0.59~0.63

In this study, the fine-scale soybean distribution information provided by Planet images was taken as the ground truth, in order to perform accuracy assessment of the results obtained from GF-6 WFV data. Planet is the world's largest constellation of microsatellites, consisting of hundreds of Dove satellites, which can provide 4- or 8-band multispectral images with a spatial resolution of 3–5 meters, and the data can be updated daily. In this study, 4-band (B, G, R, NIR) Planet images were obtained through the Planet Explorer website (https://www.planet.com/explorer/, accessed on 22 February 2019). The imaging dates were 16, 18 and 19 August (covering the Guoyang study area) and 17–19 August (covering the Mingguang study area). The images were L3B product with a spatial resolution of 3 m, which had undergone orthorectification and radiometric correction, and were stored in TIFF format.

To gain a comprehensive understanding of the composition of land cover types within the study areas and collect valuable sample data, field surveys were conducted from 28 to 31 August and from 7 to 9 September 2019. During the field campaigns, a handheld GPS device (Trimble Geo7X, Trimble Navigation Ltd., Sunnyvale, USA) was used to measure the coordinates of representative survey plots, and the corresponding land cover types were also recorded. A total of 458 samples were collected in Guoyang, they were categorized into soybean, corn, sorghum, sweet potato, and others (including sesame and vegetables), with sample size of 191, 166, 38, 24, and 39, respectively. In Mingguang, a total of 528 samples were collected, including 136 soybean, 103 corn, 217 rice, 21 peanut, 23 sweet potato, 12 sorghum, and 16 other types (mainly mugwort and stevia).

In addition, this study also utilized the 10-m resolution global land cover product, i.e., FROM-GLC10 released by Tsinghua University (http://data.starcloud.pcl.ac.cn/zh/resource/1, accessed on 22 February 2019). This product offers spatial distribution of 10 major land cover types, including information on cropland distribution (coded as 10), it was used to eliminate non-agricultural pixels within the study areas.

2.3. Methods

2.3.1. Preprocessing of GF-6 WFV Image

The acquired WFV image is in L1A level and requires preprocessing such as orthorectification, radiometric calibration, atmospheric correction, and clipping. In this study, the ENVI 5.3 software was used to perform these operations. For orthorectification,

NASA-DEM data with a spatial resolution of 30 m available from the LP-DAAC Data Pool (https://lpdaac.usgs.gov/products/nasadem_hgtv001/, accessed on 22 February 2019) was applied and the individual tiles needed to be mosaiced together in advance with the support of ArcGIS 10.7. Radiometric calibration coefficients for the GF-6 WFV sensor in 2019 was obtained from the China Centre for Resources Satellite Data and Application (https://www.cresda.com/zgzywxyyzx/zlxz/list/zlxz_1.html, accessed on 22 February 2019), and atmospheric correction was conducted employing the FLAASH module. Finally, the image was subset using the administrative boundary of Guoyang and Mingguang to obtain the data within these study areas. It was worth noting that even after orthorectification, there still was a certain geolocation bias in the WFV image. In our research, a spatial registration was performed adopting two scenes of Landsat 8-OLI image to correct the geolocation deviation, with the results of image sharpening (15-m spatial resolution, NNDiffuse method) involving panchromatic and multispectral bands serving as reference data. The reference OLI image for Guoyang was acquired on 27 September 2019 (Path/Row: 122/037), and that for Mingguang was acquired on 19 August (Path/Row: 121/037).

2.3.2. Elimination of Non-Cropland Objects

This study employed a stepwise hierarchical extraction strategy. Firstly, a set of filtering rules was established to remove non-cropland pixels corresponding to water body, artificial objects, sparse trees, and bare soil from the image. This was done to eliminate interference from non-agricultural cover types and limit the targets of interest to field vegetation, which were the focus of the subsequent feature selection procedure. For GF-6 WFV data, the most easily distinguishable water bodies were first removed with *NDWI* (Normalized Difference Water Index) [30] to achieve the separation of land pixels, by applying appropriate thresholds. For built-up areas, since the band configuration of GF-6 WFV does not support the calculation of indices like NDBI (Normalized Difference Built-up Index) [31], a threshold-based segmentation using NDVI (Normalized Difference Vegetation Index) was implemented to exclude buildings and other man-made objects along with bare soil. Finally, a reflectance threshold of near-infrared band (B4) was applied to eliminate remaining sparse trees and other non-cropland cover types.

$$NDWI = \frac{\rho_{green} - \rho_{nir}}{\rho_{green} + \rho_{nir}} \quad (1)$$

In the above formula, ρ_{green} and ρ_{nir} represent the reflectance of the green band (B2) and the near-infrared band (B4) respectively.

In this study, the spatial distribution of cropland (code ID = 10) provided by the FROM-GLC10-2017 land cover product [32] was adopted to trim the resulting data generated from the abovementioned steps, which served as the final filtering rule for the decision tree, so as to further remove non-cropland pixels that might exist in the preliminary results. To match the pixel size of the GF-6 WFV image, the spatial resolution of the cropland layer was resampled to 16 m using the nearest neighbor method. In consequence, a mask was generated to process the image within the study areas and obtain the overall distribution of field vegetation, which would be used for subsequent soybean mapping.

2.3.3. Generation of Additional Feature Variables

Generation of Additional Feature Variables

By introducing several additional features e.g., multiple vegetation indices, color space components and texture features derived from the spectral bands, more information can be extracted from the original image, which is conductive to enhancing its performance in crop mapping [19,21,33]. With the support of ENVI 5.3 software, nine vegetation indices (as shown in Table 2), three color space components (H, S, V), and eight texture features were generated based on the WFV data, among which the vegetation indices were calculated using the Bandmath tool. Taking any single band as input, eight texture features could be derived through Gray-Level Co-occurrence Matrix (GLCM), since the original

image contains eight spectral bands, calculating texture features for each band would result in excessively large output data. Therefore, the WFV image was first subjected to PCA transformation, and the first principal component (PCA1) with the highest amount of information was selected to produce the texture features (using a 3 × 3 window). Subsequently, visual interpretation was applied to these textures for preliminary selection, and only four texture features (viz., Mean, Entropy, Second Moment and Homogeneity), showing relatively distinct soybean field signatures, were retained as candidates.

Table 2. The vegetation indices adopted in this study.

Vegetation Index	Expression	Reference
Enhanced Vegetation Index (EVI)	$\frac{2.5\times(\rho_{nir}-\rho_{red})}{\rho_{nir}+6\times\rho_{red}-7.5\times\rho_{blue}+1}$	[34]
Soil-Adjusted Vegetation Index (SAVI)	$\frac{1.5\times(\rho_{nir}-\rho_{red})}{\rho_{nir}+\rho_{red}+0.5}$	[35]
Normalized Difference Vegetation Index (NDVI)	$\frac{\rho_{nir}-\rho_{red}}{\rho_{nir}+\rho_{red}}$	[36]
Red Edge NDVI-1 ($NDVI_{re1}$)	$\frac{\rho_{nir}-\rho_{re1}}{\rho_{nir}+\rho_{re1}}$	[37]
Red Edge NDVI-2 ($NDVI_{re2}$)	$\frac{\rho_{nir}-\rho_{re2}}{\rho_{nir}+\rho_{re2}}$	[37]
Green Normalized Difference Vegetation Index (GNDVI)	$\frac{\rho_{nir}-\rho_{green}}{\rho_{nir}+\rho_{green}}$	[38]
Red Edge Normalized Difference Vegetation Index (NDre)	$\frac{\rho_{re2}-\rho_{re1}}{\rho_{re2}+\rho_{re1}}$	[39,40]
Green Chlorophyll Index (CI_{green})	$\frac{\rho_{nir}}{\rho_{green}}-1$	[41,42]
Wide Dynamic Range Vegetation Index (WDRVI)	$\frac{0.2\times\rho_{nir}-\rho_{red}}{0.2\times\rho_{nir}+\rho_{red}}$	[43]

Note: ρ represents the reflectance of specified band, i.e., blue refers to the blue band, green refers to the green band, red refers to the red band, and nir refers to the near-infrared band; re1 represents the red edge-1 band (B5), and re2 represents the red edge-2 band (B6).

We adopted color space components, i.e., H, S, and V as additional candidate features in this study. The HSV color space was derived from the RGB space, it was necessary to select three bands which were then assigned to the R (red), G (green), and B (blue) color channels. In order to select the three bands with rich spatial details and minimal information redundancy for color space transformation (RGB to HSV), this study employed the *OIF* (Optimum Index Factor) as an evaluation criterion to screen out the optimal combination of bands [44]. A higher *OIF* value indicates lower correlation between the bands (less information redundancy) and more detailed information about ground objects, which is conducive to revealing differences between different land cover types. *OIF* is calculated as follows:

$$OIF = \sum_{i=1}^{3} S_i / \sum_{i,j=1}^{3} R_{ij} \quad (2)$$

In the formula, "S_i" represents the standard deviation of the *i*-th band, and "R_{ij}" represents the correlation coefficient between the *i*-th and *j*-th bands.

The original 8 bands of GF-6 WFV data were organized into various combinations, with each consisting of 3 bands, generating a total of 56 combinations. The *OIF* of each combination was calculated and arranged in descending order, as a result, the optimal band combination for Guoyang was determined to be bands 3, 4, and 6, while for Mingguang, the best combination was identified as bands 4, 6, and 8 (as shown in Table 3). By assigning different bands to the R, G, and B channels for color composite, it was found that designating band 3 as R, band 4 as G, and band 6 as B achieved the best result for the Guoyang image. Similarly, for the Mingguang image, assigning band 4 as R, band 6 as G, and band 8 as B attained the optimal effect. These band combinations were then subjected to HSV transformation to obtain three color space components (Hue, Sat, Val).

Table 3. *OIF* values corresponding to various band combination forms of GF-6 WFV data (top 8).

Order	Guoyang		Mingguang	
	Band Combination	***OIF***	**Band Combination**	***OIF***
1	3-4-6	0.1268	4-6-8	0.1124
2	4-6-8	0.1261	3-4-6	0.1092
3	1-4-6	0.1234	4-6-7	0.1090
4	4-6-7	0.1231	1-4-6	0.1085
5	2-4-6	0.1202	2-4-6	0.1077
6	4-5-6	0.1175	4-5-6	0.1044
7	3-4-8	0.0957	3-4-8	0.0863
8	1-3-4	0.0933	2-4-8	0.0856

2.3.4. Feature Importance Evaluation

There were a total of 24 candidate features, this study adopted a feature selection strategy to filter out a subset of features sensitive to soybean cropland for its mapping. The first step was to evaluate the importance of each candidate feature and rank them accordingly. The ReliefF algorithm is a widely accepted and well-performing method for feature importance evaluation [45]. In this study, based on the WFV images after removing non-cropland pixels, typical samples for different crop types in Guoyang and Mingguang were collected. The samples from Guoyang included soybean, corn, sorghum, and other vegetation (mainly vegetables and sweet potato), while from Mingguang included rice, soybean, corn, peanut, and other vegetation (as defined above). The ReliefF algorithm was executed in MATLAB R2018a software to evaluate the weights of candidate features. Using the sample datasets of field vegetation, the output results were saved in .csv format. To reduce the potential uncertainty arising from the randomness of the ReliefF algorithm in feature weight evaluation, the algorithm was executed 20 times, and the average of the results from 20 runs was taken as the final weight for each feature. This process enabled the determination of the importance ranking of the candidate features.

2.3.5. Classifier-Specific Feature Selection

Three machine learning algorithms, i.e., Support Vector Machine (SVM), Random Forest (RF), and Back-Propagation Neural Network (BPNN) were employed for feature selection and soybean fields extraction. The feature selection process involved two crucial aspects: determining the optimal feature dimension (i.e., the number of features that achieved the highest accuracy) and deciding which features to retain or discard (as certain features might lead to reduced accuracy). Based on the importance ranking of candidate features, this study adopted a Sequential Forward Selection (SFS) method to determine the optimal dimension of the feature subset, which was performed coupled with specified classifier, taking into account the unique characteristics of different algorithms. In this approach, the feature with the highest weight was firstly input into certain classifier to obtain the initial classification accuracy. Subsequently, the next feature with a slightly lower weight was incrementally added, forming new data in combination with the previously input features, and the corresponding overall accuracy (OA) was then calculated. Features were added one at a time based on the weight ranking for accuracy evaluation, until all candidate features had been inputted. Finally, the corresponding curve depicting the variation in overall accuracy was generated.

For a specified feature, if its inclusion had a negative impact on classification accuracy, i.e., resulting in a decrease in accuracy or remaining unchanged, it would be discarded. The well-behaved features for each classifier could be singled out according to the accuracy curve, and the resultant feature-subsets corresponding to different classifiers were used as input data for classification to extract soybean planting areas. In this study, the process of feature selection was performed in MATLAB R2018a software and carried out independently for the two study areas.

The Random Forest (RF) algorithm supports high-dimensional data, has a high tolerance for noise and offers accurate predictions, making it widely applicable. RF algorithm has two key parameters: the number of variables randomly chosen for decision tree splits (Mtry), and the total number of trees generated during model execution (Ntree). Some studies have shown that the RF algorithm can achieve satisfactory results with default parameters [46]. In this study, Mtry was set to the square root of the total number of features participating in classification, and Ntree was set to 100. The Back-Propagation Neural Network (BPNN) possesses strong nonlinear mapping capabilities and is effective for addressing real-world nonlinear modeling problems. For the BPNN, this study used a single hidden layer network with 1000 iterations and a learning rate of 0.02, while keeping other parameters at their default settings. The Support Vector Machine (SVM) algorithm distinguishes between different land cover types by constructing an optimal separating hyperplane. Previous studies have demonstrated that among the commonly used four kernel functions (i.e., polynomial, linear, sigmoid, and RBF), the RBF kernel function is more suitable for distinguishing between different crop types [47]. Therefore, in this study, the RBF kernel function was applied in the SVM classification model for soybean identification. Additionally, Gamma was set to the reciprocal of the number of input features, and the classification threshold was set to 0 to ensure that each pixel would be defined, and other parameters were kept at their default values.

2.3.6. Model Selection and Its Performance Evaluation

The three candidate models (i.e., ReliefF-RF, ReliefF-SVM, and ReliefF-BPNN) were deduced to exhibit distinct behaviors in soybean identification, allowing for the selection of the best-performing one among them. With the support of ENVI 5.3 platform, the three machine learning algorithms were individually executed to perform classification using the feature-subsets filtered out in the previous step as input data, acquiring the distribution of soybean planting areas corresponding to each model. In order to examine the extraction effects of different models and achieve model selection, this study established two sets of ground samples in the two study areas, each was 313 × 313 pixels (about 5 km of side length) in size, with eight ground samples placed in Guoyang and six in Mingguang. The layout of these samples took into account the composition of land cover types and variations in soybean proportions (as shown in Figure 3). The Planet images within each ground sample were classified using the Maximum Likelihood method, and the classification results were aggregated into two categories: soybean and non-soybean, which served as ground truth to evaluate the extraction accuracy of GF-6 WFV data.

With the support of ground truth data, a confusion matrix was generated to evaluate the performance of each model, and then four accuracy evaluation metrics derived from the confusion matrix were adopted, i.e., Overall Accuracy (OA), Producer's Accuracy (PA), User's Accuracy (UA), and Kappa coefficient. The Kappa coefficient makes full use of the information in the confusion matrix and provides a comprehensive measure of classification accuracy. In this study, the 3-m resolution Planet images were classified using the Maximum Likelihood method to obtain fine-scale soybean distribution maps for the existing ground samples. These results were then resampled using the nearest neighbor method to match the 16-m resolution of GF-6 WFV data, thereby serving as the ground truth. Kappa coefficients were subsequently calculated to evaluate the accuracy of extraction results obtained from different models, facilitating model selection.

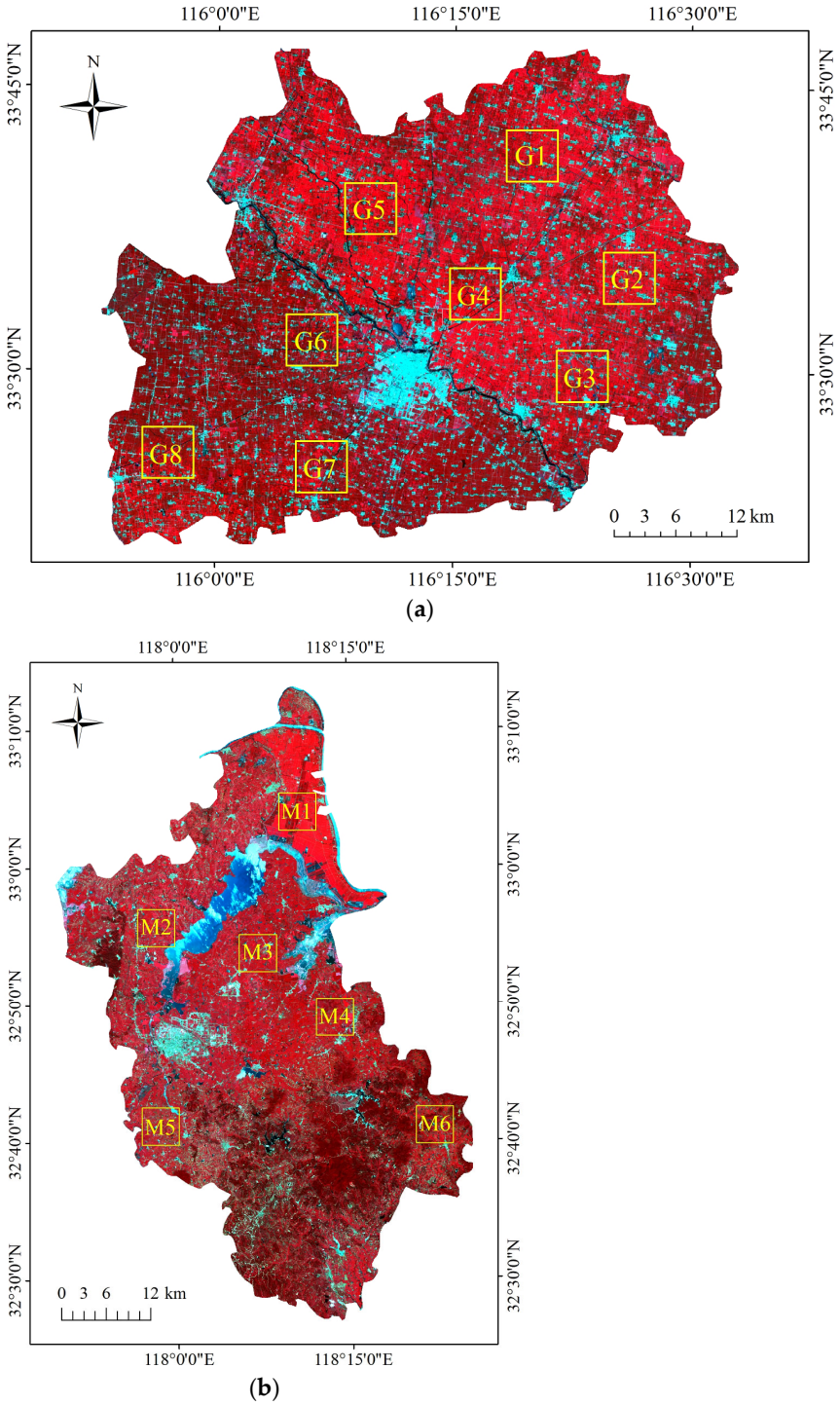


Figure 3. Spatial distribution of the labeled ground samples established in the two study areas. Note that the text in yellow indicates the number for each ground sample, where 'G' represents the Guoyang study area, and 'M' represents the Mingguang study area; the base maps are GF-6 WFV data acquired on 16 August 2019 in false-color (R(4): G(3): B(2)) composites. (**a**) Guoyang. (**b**) Mingguang.

This article compared the selected feature-subset with two alternative input data forms in terms of the extraction effect, aiming to examine the performance of the optimum model in soybean mapping. Three schemes were designed in this section, scheme A involved the original spectral bands, while scheme B adopted total candidate features without feature selection, and scheme C employed the selected features obtained in the previous section (Table 4). With the aid of the same training samples and ground truth data, different forms

of feature dataset were inputted into the best-performing classifier to derive mapping accuracy corresponding to different schemes. It could reveal whether the selected features had an advantage over the other two forms of input data in soybean identification, thus evaluating the practical significance of the feature selection operation.

Table 4. Different extraction schemes for soybean planting areas.

Scheme	Feature Combination
A	Original 8 WFV spectral bands
B	24 total candidate features
C	Selected feature subset

This study mainly involved the following 5 steps: (1) preprocessing the WFV image; (2) removing non-cropland pixels; (3) generating the complete set of candidate features and evaluating feature importance; (4) selecting features coupled with classifiers; and (5) selecting the optimum model and evaluating its performance. Figure 4 shows the overall technical roadmap:

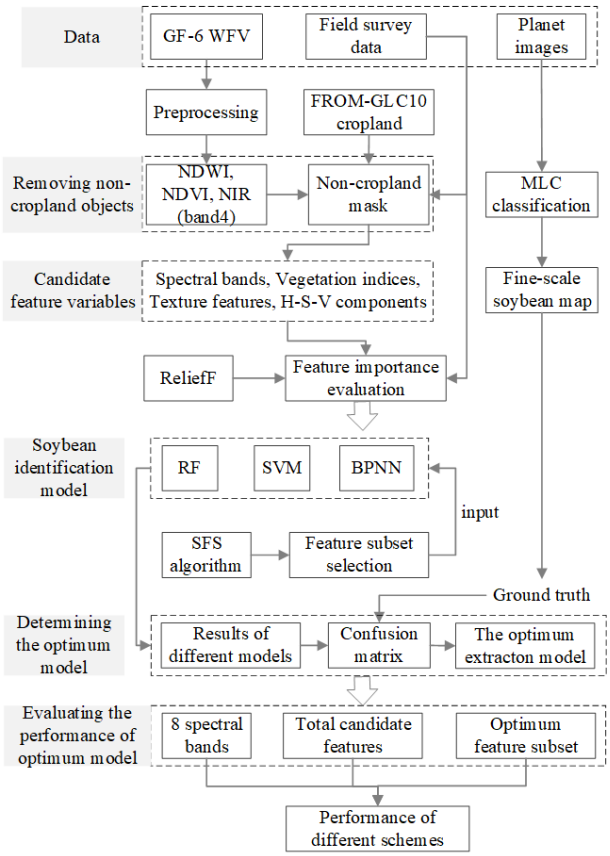


Figure 4. The technical roadmap of this study.

3. Results and Analysis

3.1. Filtering Rules for Eliminating Non-Cropland Objects

For the study areas of Guoyang and Mingguang, this study established a set of decision tree filtering rules based on *NDWI*, NDVI, and NIR band (B4) reflectance to remove non-agricultural types from the GF-6 WFV image. Additionally, the results were further revised using cropland distribution layer provided by the FROM-GLC10 product, and the target objects for subsequent work could thereby be focused on field vegetation. The filtering rules established are shown in Figure 5:

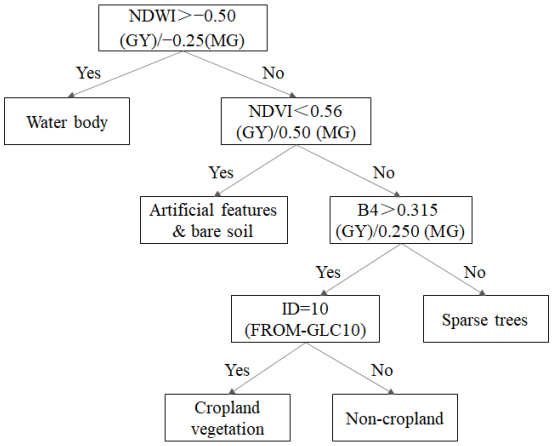
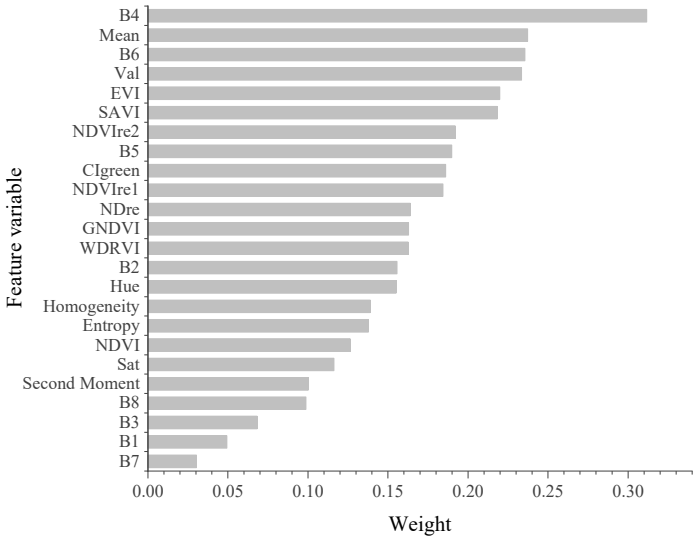


Figure 5. Decision tree filtering rules for the removal of non-cropland objects. Note that GY represents Guoyang and MG denotes Mingguang, respectively; and B4 is the reflectance of the near-infrared band.

3.2. The Importance and Ranking of Candidate Features

Based on the GF-6 WFV image after removing non-cropland pixels, sample points covering major land cover types were selected to create region of interest. The numerical values of candidate features were extracted, and the ReliefF algorithm was applied to evaluate the importance of 24 candidate feature variables in soybean identification. Results showed that for the Guoyang study area, the NIR band with a central wavelength of 830 nm (B4) ranked first, indicating that this feature contributed the most to soybean identification. The 'Mean' texture feature ranked second and the red edge-2 band (B6) ranked third, both of which were also conductive to realizing soybean-field extraction. In addition, EVI and SAVI were significant for soybean mapping compared to other vegetation indices (Figure 6a). For the Mingguang study area, the most prominent spectral features were the NIR band (B4) and red edge-2 band (B6), and the 'Mean' texture feature also demonstrated advantages. Among all vegetation indices, EVI, SAVI, and CI_{green} ranked higher (Figure 6b), which was consistent with the evaluation results obtained in the Guoyang study area.



(**a**)

Figure 6. *Cont.*

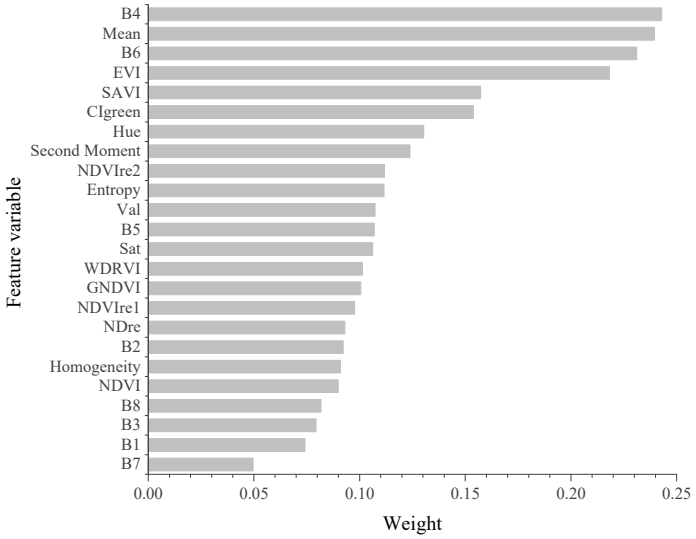


Figure 6. Ranking of the importance of each candidate feature in descending order. (**a**) Guoyang study area. (**b**) Mingguang study area.

3.3. The Selected Feature-Subset Corresponding to Different Classifiers

This study used the SFS method to realize feature-subset selection, which was conducted coupled with different classifiers. As the number of input features increased, the overall classification accuracy generally showed an increasing trend (except for SVM). For the Guoyang study area, the ReliefF-RF model reached the highest accuracy at a feature dimension of 11 (marked with red triangle in Figure 7a), and the accuracy remained stable with minimal fluctuations as more features were added. When all 24 feature factors were included in the classification, there was no significant improvement in accuracy, therefore the features that did not contribute to improve classification accuracy were discarded in this study. In consequence, the ReliefF-RF model selected the top 8 feature variables with the highest weight as the optimum feature-subset for this classifier. Similarly, the ReliefF-BPNN model's optimal feature dimension was 9, and one feature that did not improve classification accuracy was discarded (Figure 7c); the ReliefF-SVM model had the best feature dimension of 9, with 7 well-behaved features retained (Figure 7e).

Figure 7b,d,f illustrate the relationship between feature dimension and classification accuracy, as well as the determined dimensions for different classifiers in Mingguang. The RF classifier reached local maximum accuracy at a feature dimension of 9 (marked with red triangle in Figure 7b), and there was no significant change in accuracy afterwards, with only minor fluctuations within a small range. When the first 18 features were used as input, the accuracy reached its global maximum but was only 0.69% higher than that achieved with the first 9 features. Therefore, the ReliefF-RF model selected 6 out of the first 9 features as the optimum feature subset. Similarly, the ReliefF-BPNN model had the optimal feature dimension of 7, with 6 features selected; the ReliefF-SVM model also identified 6 optimum features (Figure 7f).

Due to the three machine learning algorithms differ in principle and architecture, ReliefF-RF, ReliefF-BPNN, and ReliefF-SVM models screened out different combinations of features. Table 5 provides a detailed list of the selected feature subsets for the three models, clearly demonstrating their differences.

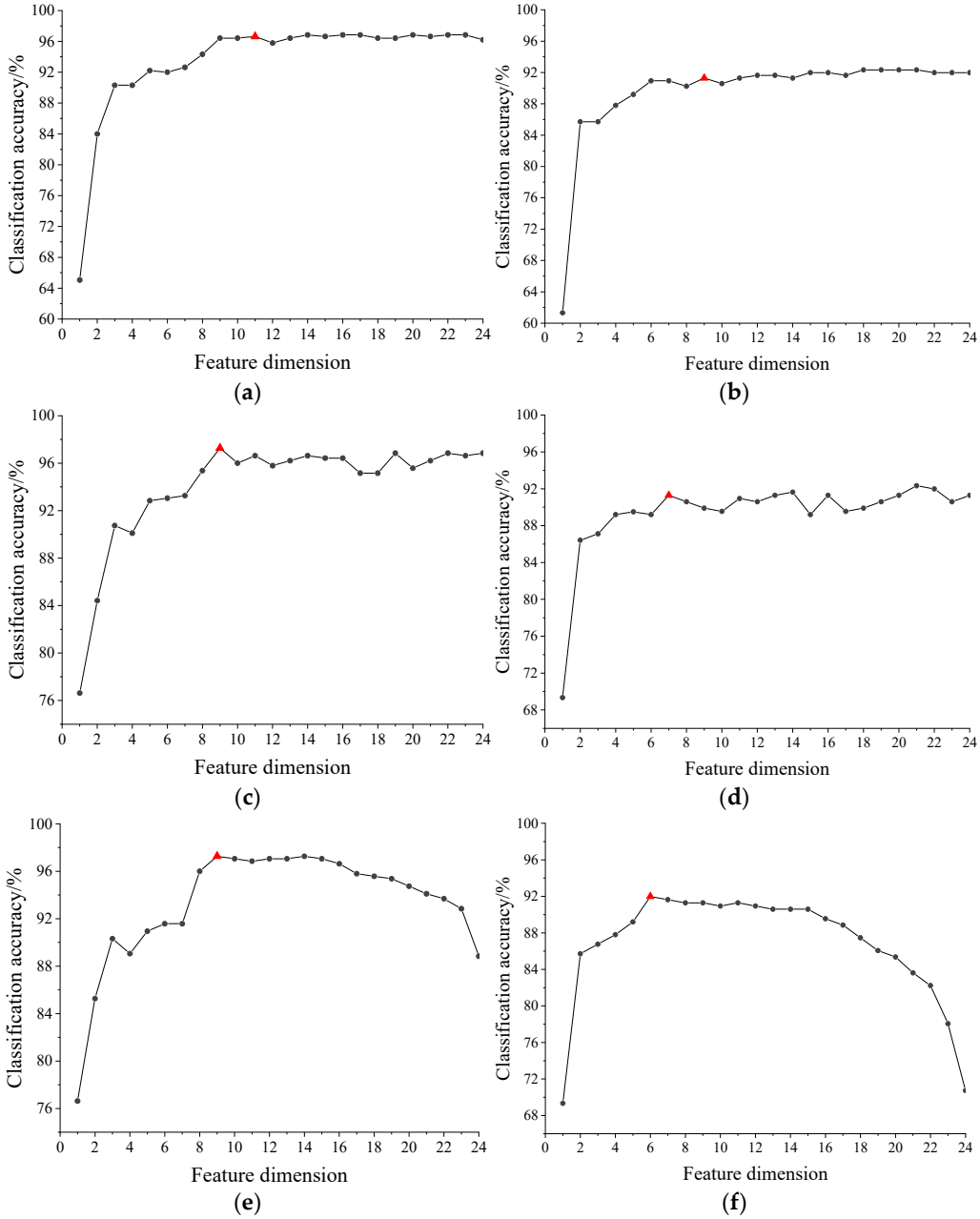


Figure 7. Relationship between classification accuracy of each model and number of input features: (**a**,**b**) ReliefF-RF; (**c**,**d**) ReliefF-BPNN; (**e**,**f**) ReliefF-SVM. Note: The optimal feature dimensions are represented by red triangles (△); (**a**,**c**,**e**) correspond to the results from Guoyang, whereas (**b**,**d**,**f**) represent the results from Mingguang.

Table 5. Optimum subset of candidate features under different classifiers.

Model	Results for Guoyang	Results for Mingguang
ReliefF-RF	B4, Mean, B6, EVI, $NDVI_{re2}$, B5, CI_{green}, NDre	B4, Mean, EVI, SAVI, CI_{green}, $NDVI_{re2}$
ReliefF-BPNN	B4, Mean, B6, EVI, SAVI, $NDVI_{re2}$, B5, CI_{green}	B4, Mean, B6, EVI, CI_{green}, Hue
ReliefF-SVM	B4, Mean, B6, EVI, SAVI, B5, CI_{green}	B4, Mean, B6, EVI, SAVI, CI_{green}

3.4. The Optimum Model for Soybean Mapping

The performance of different classifiers in soybean identification was evaluated based on the detailed distribution of soybean fields within each ground sample, which was derived from Planet images. For the two study areas, i.e., Guoyang and Mingguang, this study took the selected feature-subset corresponding to each classifier as input to perform classification and obtained soybean planting areas. Table 6 (G1~G8) presents the extraction accuracy of the GF-6 WFV image in Guoyang using the three models. In comparison to the RF and BPNN classifiers, the ReliefF-SVM model generally exhibited higher extraction accuracy (except for sample 5) and outperformed the other two models, indicating the results were closer to the actual soybean maps. Therefore, this study selected the ReliefF-SVM model for the extraction of soybean planting areas in Guoyang.

Table 6. The extraction accuracy of different models in the study areas of Guoyang and Mingguang.

Ground Sample	ReliefF-RF				ReliefF-BPNN				ReliefF-SVM			
	PA/%	UA/%	OA/%	Kappa	PA/%	UA/%	OA/%	Kappa	PA/%	UA/%	OA/%	Kappa
G1	85.00	80.64	88.39	0.74	82.16	74.45	84.91	0.67	90.20	78.01	88.45	0.75
G2	89.09	70.86	76.19	0.68	86.38	66.35	83.32	0.63	92.05	68.53	85.41	0.69
G3	84.59	75.03	89.93	0.73	83.47	71.14	88.35	0.69	88.41	73.31	89.88	0.73
G4	91.33	79.50	89.33	0.77	90.47	74.30	86.48	0.71	93.88	78.75	89.58	0.78
G5	93.99	73.97	87.24	0.72	91.54	69.83	84.33	0.67	96.13	71.32	86.12	0.71
G6	79.32	82.23	95.15	0.66	53.48	80.50	94.60	0.61	65.54	80.62	95.44	0.70
G7	74.49	84.71	94.98	0.76	67.68	85.37	94.34	0.72	78.88	86.15	95.64	0.80
G8	59.27	85.16	93.03	0.66	60.25	80.25	93.74	0.65	70.27	78.19	94.35	0.71
M1	94.30	82.38	89.16	0.78	95.39	79.79	87.27	0.75	91.39	83.13	88.00	0.76
M2	66.39	71.82	87.85	0.61	85.06	74.65	91.07	0.74	69.49	87.66	91.79	0.73
M3	63.58	68.75	94.89	0.63	76.62	72.94	95.94	0.72	70.71	80.60	96.37	0.73
M4	64.94	65.75	92.59	0.61	81.09	67.68	93.80	0.70	67.60	81.58	94.87	0.71
M5	67.94	70.38	95.03	0.66	92.36	58.03	93.90	0.68	61.27	93.85	96.50	0.72
M6	71.83	67.26	96.93	0.68	84.38	56.89	96.12	0.66	75.80	73.19	97.47	0.73

Note: OA—Overall Accuracy, PA—Producer's Accuracy, UA—User's Accuracy.

Table 6 (M1~M6) presents the extraction accuracy of the three models within each ground sample in Mingguang. It could be observed that the ReliefF-SVM model generally outperformed the other two models in soybean mapping. It is worth noting that sample 1 located in the northern part has a flatter terrain and a larger-scale, more concentrated soybean cultivation. Therefore, all three models yielded relatively ideal extraction results in this sample. By contrast, the terrain in other ground samples is more rugged with higher surface fragmentation, leading to lower extraction accuracy for all models compared to sample 1. The user's accuracy (UA) of ReliefF-RF and ReliefF-BPNN models was lower than that of the ReliefF-SVM model in all ground samples, indicating that these two models misclassified more non-soybean objects as soybean. In comparison to the ReliefF-RF and ReliefF-BPNN models, the ReliefF-SVM model demonstrated more stable performance and relatively higher accuracy. Therefore, this study selected the ReliefF-SVM model for soybean mapping in the Mingguang study area.

3.5. Performance of the Optimum Model in Soybean Mapping

Based on the SVM algorithm and different forms of input datasets, soybean distribution maps within two study areas were generated, and the performance of the optimum model, i.e., ReliefF-SVM in soybean identification was examined on the basis of accuracy level (Table 7). For the Guoyang study area, the soybean map derived from ReliefF-SVM within each ground sample was highly consistent with the ground truth (Figure 8). Scheme C based on the selected feature subset achieved Kappa coefficients above 0.69 for all 8 ground samples. In comparison to the extraction results using the 8 spectral bands (Scheme A), the Kappa coefficients were improved by at least 0.02 excepting for sample

8. Scheme B which used the total features as input, had a Kappa coefficient that was 0.05 higher than Scheme C on sample 8, but showed no advantage on the remaining samples. However, Scheme C reduced the data volume by 70.83% compared with Scheme B, while ensuring accuracy. Therefore, the selected feature-subset not only maintained accuracy but also significantly reduced data volume and computational cost, demonstrating its advantage in soybean identification.

Table 7. Comparison of the accuracy of different schemes based on GF-6 WFV data.

Study Area	Scheme	Kappa							
		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
Guoyang	A	0.71	0.61	0.68	0.75	0.66	0.66	0.78	0.75
	B	0.70	0.68	0.73	0.75	0.71	0.70	0.78	0.76
	C	0.75	0.69	0.73	0.78	0.71	0.70	0.80	0.71
Mingguang	A	0.76	0.70	0.70	0.70	0.71	0.70	-	-
	B	0.74	0.74	0.68	0.72	0.73	0.66	-	-
	C	0.76	0.73	0.73	0.71	0.72	0.73	-	-

For the Mingguang study area, the distribution of soybean planting areas within each ground sample also showed a high degree of consistency with the reference map (Figure 9). Results of different schemes showed that, when compared to Scheme A (employing 8 spectral bands), the ReliefF-SVM model based on the selected features (Scheme C) exhibited an identical Kappa coefficient in sample 1. Notably, it demonstrated an improvement of 0.03 in samples 2, 3, and 6, and a marginal increase of 0.01 in samples 4 and 5. All of the above indicated that the selected features improved or maintained extraction accuracy in all ground samples. In addition, the Kappa coefficients derived from Scheme C were slightly lower in samples 2, 4, and 5 (with a difference of 0.01), but higher in other ground samples when compared to the results obtained using all 24 features as input (Scheme B). Furthermore, the selected feature subset reduced the data volume by 75% compared to total features, thereby showcasing a distinct advantage in both extraction accuracy and data reduction.

Figure 8. *Cont.*

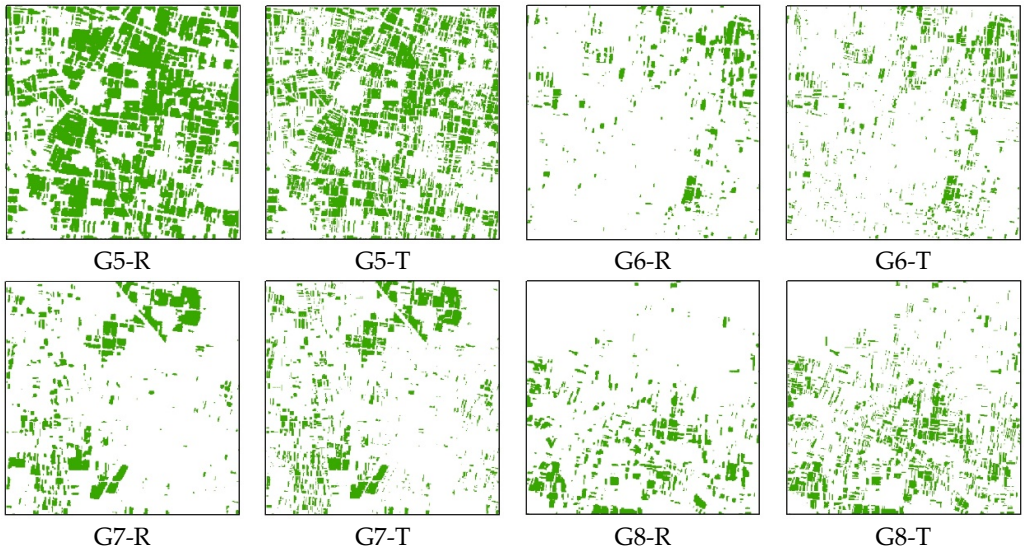


Figure 8. The extraction results of ReliefF-SVM model and the corresponding ground truth in those ground samples in Guoyang. Note that 'R' represents the result, and 'T' indicates the ground truth; the soybean planting areas are shown in green.

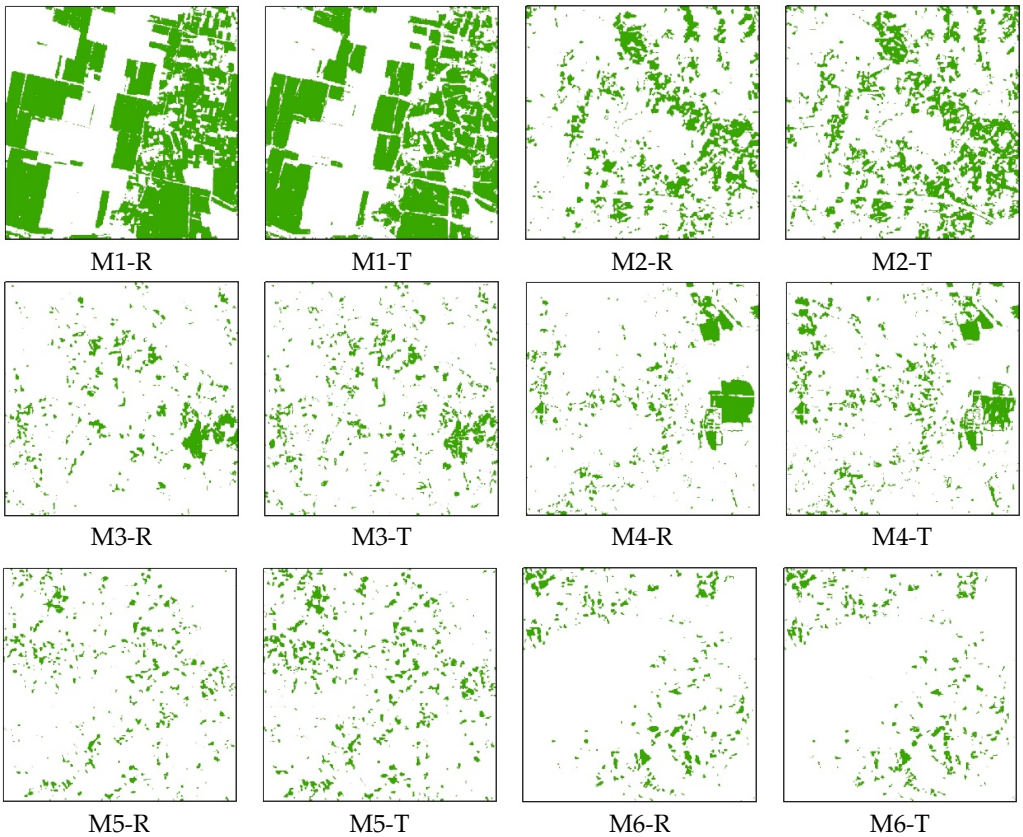
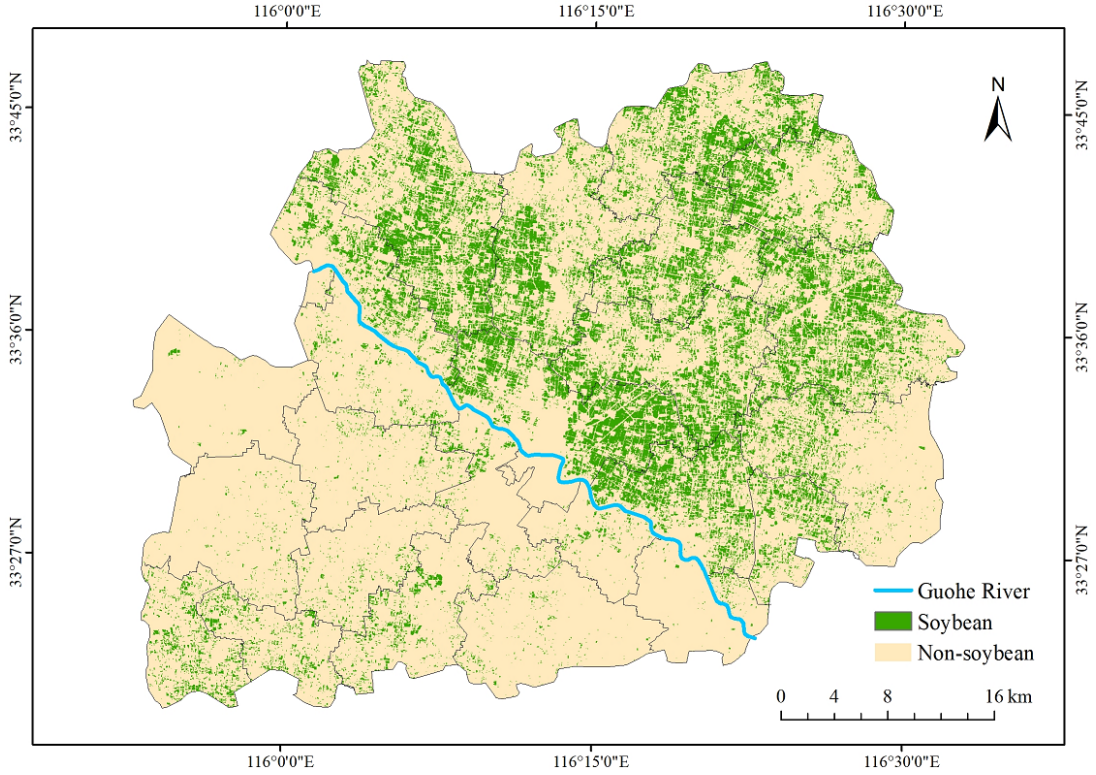


Figure 9. The extraction results of Relief-SVM model and the corresponding ground truth in those ground samples in Mingguang. Note that 'R' represents the result, and 'T' indicates the ground truth; the soybean planting areas are shown in green.

Using the SVM classifier and the selected feature subset as input, the soybean planting areas in Guoyang and Mingguang were extracted. The soybean fields in Guoyang were mainly distributed in the region north of the Guohe River, which dominated among summer crops; the region south of the Guohe River had a smaller cultivation scale and a more scattered distribution (Figure 10a). In addition, the prevalence of smallholder farming practices resulted in widespread mixed cropping of soybean and other summer crops, there were few large-size contiguous soybean fields. The scale of soybean cultivation in Mingguang was significantly smaller than that in Guoyang. The soybean fields were mainly concentrated in the northeastern part and the north bank of Nushan Lake, where several large farms were located. The cropland here had been leveled and modified to facilitate mechanized farming, making it more regular and equipped with agricultural infrastructure, hence large-scale contiguous planting was the main pattern in this area. In other areas of Mingguang, the distribution of soybean fields was highly dispersed, especially in the southern mountainous region where only a few isolated soybean fields could be observed (Figure 10b).



(**a**)

Figure 10. *Cont.*

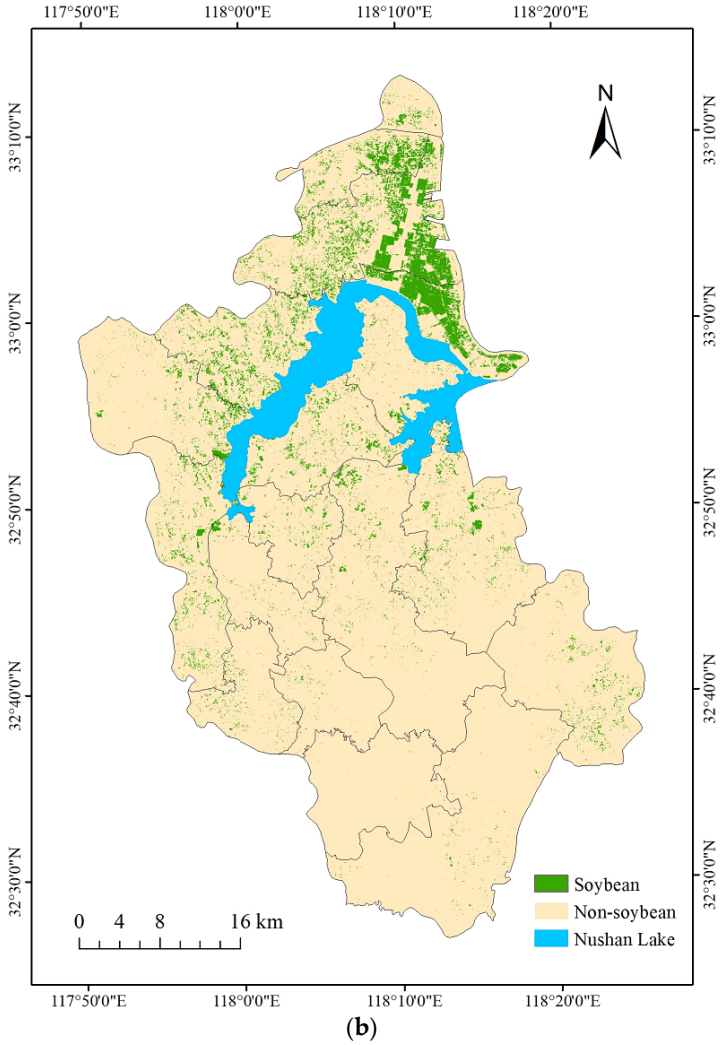


Figure 10. Spatial distribution of soybean-cultivated areas in Guoyang and Mingguang during the 2019 growing season. (**a**) Guoyang. (**b**) Mingguang.

4. Discussion

The Chinese GF-6 WFV sensor is known for its two red-edge bands, up to now, it has been less utilized in crop identification research. We selected two study areas with markedly different natural environments and planting structures, belonging respectively to the Huang-Huai-Hai Basin and the Yangtze River Basin, to explore its application in soybean extraction. Our research enriched the case studies on soybean mapping under similar cropping conditions, and the method presented in this article may serve as a valuable reference for subsequent research in this direction. Moreover, the detailed soybean distribution map provides an objective and reliable basis for local agricultural departments to carry out agricultural monitoring and growth assessment, and then help them to formulate reasonable policies to promote the development of soybean industry.

The study areas are situated in the transitional zone between northern and southern China, where the weather conditions are variable. In light of the fact that the soybean growing season coincides with the hot and rainy summer, the high frequency of cloud cover significantly limits the availability of usable optical imagery. Therefore, utilizing single-phase image for soybean mapping is more advantageous and suitable for the actual situation of the study areas. In light of the significant differences in the scale or proportion of water bodies, built-up areas, trees, bare soil, and other cover types in different regions, this study adopted a stepwise hierarchical extraction strategy. By establishing a set of filtering rules, non-cropland cover types mentioned above were eliminated to minimize

their interference with the subsequent results. Follow-up work could focus on field vegetation, making the selected features more targeted and promoting the applicability and generalizability of the constructed soybean identification model across different regions.

This study involved an operation, i.e., feature selection coupled with classifier to enhance the performance of machine learning models. The goal was to fully explore the potential of the algorithm in soybean identification. For GF-6 WFV image, the ReliefF-SVM model's performance in the two study areas was significantly better than those of the other two models, indicating that the SVM classifier was suitable for soybean mapping using this data source. The optimum features selected for Guoyang and Mingguang showed slight differences. In general, the near-infrared band (B4) and red edge-2 band (B6) were found to be more sensitive to soybean, which was consistent with the findings in Tian et al. [12] and Ren et al. [20]. 'Mean' texture feature was favorable for soybean extraction, and EVI, SAVI as well as CI_{green} outperformed other vegetation indices in soybean identification. The color space components, i.e., Hue, Sat and Val did not exhibit superiority in this task. Feature selection operation helped filter out the features that significantly contributed to soybean recognition and effectively remove less important features, enhancing the execution efficiency while ensuring extraction results. Therefore, it demonstrated a promising application prospect in soybean mapping.

Our research selected Guoyang in the northern plain of Anhui Province and Mingguang in the Jianghuai hilly region as the study areas, the aim was to investigate the effectiveness and applicability of soybean identification models in regions characterized by dramatically different planting structures and diverse natural environments. The accuracy evaluation results indicated that the model after classifier and feature selection performed well in soybean identification in both two study areas. The accuracy values were acceptable, demonstrating that this approach was applicable to both the northern plain and the hilly region between the Yangtze River and the Huaihe River. Furthermore, the extraction accuracy of soybean planting areas varied in different locations. Upon scrutinizing the raw image of each ground sample, it became evident that the areas with relatively concentrated soybean fields and high pixel purity achieved high-precision soybean mapping, e.g., sample G4 in Guoyang and sample M1 in Mingguang. Even in sample G7, located in the southern part of Guoyang where soybean fields had a low proportion but were relatively concentrated with fewer instances of mixed planting, higher extraction accuracy could still be obtained. It was suggested that the proportion of soybean cultivation did not play a decisive role in extraction accuracy.

The differences in growth and soil background may lead to the phenomenon of spectral heterogeneity of soybean fields at the remote sensing level, subsequent studies will consider exploring soybean identification under different growth levels. The planting structure of crops in the study areas were complex, field vegetation that grows concurrently with soybean and can interfere with its spectral signature includes corn, sorghum, sesame, peanut, sweet potato, vegetables, and medicinal herbs, etc. The GF-6 WFV sensor has a spatial resolution of only 16 meters, leading to a significant "mixed pixel" effect at the remote sensing level, which brings great challenge to the extraction of soybean fields and results in less-than-ideal estimates for the cultivation area. Future work will consider adopting a pixel unmixing strategy to estimate soybean abundance at the pixel level, thereby obtaining more accurate scale of soybean cultivation.

This study relies solely on a single-phase WFV image for soybean mapping, which may introduce uncertainty in the results and somewhat impact the persuasiveness of the conclusions. This is primarily attributed to our ground survey activities conducted exclusively in 2019, during which we collected survey samples from two study areas. Additionally, the suboptimal coverage of imagery data during the subsequent soybean growing seasons, specifically during the pod-setting phase, and the less-than-ideal growth conditions of soybeans further contributed to this limitation. Future work involves expanding soybean mapping based on WFV imagery at a larger spatiotemporal scale, and evaluating the model's performance in different years and different regions.

Some limitations exist in current study. Due to time and manpower constraints, the survey samples utilized in this study still appeared to be insufficient. Future research will conduct more comprehensive field surveys to obtain sufficient sample data, and special attention will be given to the areas south of the Guohe River and the southern mountainous region of Mingguang. The classifiers adopted in this study are shallow learning algorithms, which have low requirements for training samples and relatively fast execution speed, however, the extraction accuracy of these algorithms may easily reach a ceiling. Future study will adopt more machine learning algorithms and the ensemble learning strategies will also be introduced into soybean mapping to comprehensively utilize the advantages of multiple machine learning algorithms. Moreover, deep learning algorithms are considered for subsequent work with the aim of further improving extraction accuracy. Last but not least, the redundancy between the selected features was not considered in the feature selection process of this study, this issue is expected to be addressed in subsequent research.

5. Conclusions

This study is conducted based on Chinese GF-6 WFV image obtained during the pod-setting stage of soybean, which has been considered to be the most critical period for achieving soybean mapping. Three machine learning algorithms were adopted to establish soybean identification model for typical soybean-producing areas in the northern plain of Anhui Province and the hilly region south of the Huaihe River. The distribution of soybean fields within the study areas during the 2019 growing season was extracted, and the results were validated based on the fine-scale soybean maps obtained from Planet images. The main conclusions are summarized as follows:

This article adopted a hierarchical extraction strategy, which first established a set of filtering rules to eliminate non-farmland objects and reduce their interference with soybean extraction results. The follow-up work could focus on field vegetation, making the selected features more targeted and improving the model's generalizability across regions. Furthermore, the applicability and implementation effect of this method in the northern plain and the Jianghuai hilly region of Anhui Province was also effectively verified.

The extraction of soybean planting areas was conducted based on the GF-6 WFV image, and the results indicated that the ReliefF-SVM model performed the best in both two study areas. Using the selected feature subset as input, the Kappa coefficients for the ground samples in Guoyang ranged from 0.69 to 0.80, while those in Mingguang fell within the range of 0.71 to 0.76. The accuracy of extraction was higher in regions where soybean fields were relatively concentrated and less mixed cropping of soybean and other crops, while the proportion of soybean cultivation did not seem to play a decisive role in the extraction results.

Feature selection can be used to filter out a subset of features that are dedicated to soybean identification. In this study, the feature selection process was performed coupled with different classifiers to single out the most discriminating features according to the accuracy they could achieve, ensuring maximum consideration of the specificity of different classification algorithms. In the two study areas situated in northern plain and Jianghuai hilly region, the near-infrared band (B4), red edge bands (B5 and B6) were found to be more sensitive to soybean. EVI, SAVI, and CI_{green} demonstrated advantages over other commonly used vegetation indices in soybean identification, and the 'Mean' texture feature also contributed to successful soybean mapping. In addition, the extraction accuracy derived from the selected features was generally higher than that obtained using the original spectral bands, and in most cases, it was also better than the extraction results when using all features as input. Moreover, it reduced data volume and computational overhead by more than 70% compared to the latter, indicating that the optimum subset of features could improve execution efficiency while ensuring accuracy.

This study established a soybean identification model based on single-phase GF-6 WFV image to adequately adapt to the reality of frequent cloud cover and limited clear observation in the study areas. Furthermore, a series of measures such as hierarchical

extraction, classifier and feature selection were taken to optimize the extraction scheme. The established model not only achieved a balance between data volume and extraction accuracy, but also was proved to be applicable across the vastly different planting structures in the northern plain and Jianghuai hilly region of Anhui Province. Moreover, the mapping accuracy could meet practical requirements. This research can to some extent compensate for the work on soybean identification in China's primary soybean-producing regions, i.e., the Huang-Huai-Hai and Yangtze River basins, where little attention has been received and there is scarcity of detailed soybean distribution maps. It also expands the scope of research case involving machine learning methods combined with feature-selection strategy to achieve reliable soybean mapping, and could provide insight into soybean identification in regions with complex planting structures.

Author Contributions: Conceptualization, B.S.; methodology, B.S.; software, M.Z.; validation, J.H. and M.Z.; formal analysis, B.S.; investigation, Q.Y.; resources, B.S.; data curation, M.Z.; writing—original draft preparation, B.S.; writing—review and editing, B.S. and J.H.; visualization, B.S. and J.H.; supervision, L.H.; project administration, L.H.; funding acquisition, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Open Research Fund of National Engineering Research Center for Agro-Ecological Big Data Analysis & Application (AE202101), and The National Key Research and Development Program of China (2019YFE0115200).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the first au-thor on request.

Acknowledgments: The authors appreciate China Centre for Resources Satellite Data and Application which provides GF-6 WFV image for this research, we also thank Planet Corporation for pro-viding high-resolution satellite images for the validation work of this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, X.; Wang, F.; Xu, D.B.; Liu, Q. Application research on the method for extracting soybean covered areas based on the pixel unmixing. *Trans. Chin. Soc. Agric. Eng.* **2008**, *24*, 213–217. (In Chinese)
2. Zhong, L.H.; Yu, L.; Li, X.C.; Hu, L.; Gong, P. Rapid corn and soybean mapping in US Corn Belt and neighboring areas. *Sci. Rep.* **2016**, *6*, 36240. [CrossRef]
3. Wang, L.M.; Liu, J.; Yang, L.B.; Yang, F.G.; Fu, C.H. Impact of short infrared wave band on identification accuracy of corn and soybean area. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 169–178. (In Chinese)
4. Liu, J.; Wang, L.M.; Teng, F.; Yang, L.B.; Gao, J.M.; Yao, B.M.; Yang, F.G. Impact of red-edge waveband of RapidEye satellite on estimation accuracy of crop planting area. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 140–148. (In Chinese)
5. Liu, J.; Wang, L.M.; Yang, F.G.; Yao, B.M.; Yang, L.B. Recognition Ability of Red Edge and Short Wave Infrared Spectrum on Maize and Soybean. *Chin. Agric. Sci. Bull.* **2018**, *34*, 120–129. (In Chinese)
6. Wang, E.M.; Li, W.G.; Gu, X.H.; Zhang, Y.H.; Chen, H. Planting area extraction of maize based on spectral features differentiation. *Jiangsu J. Agric. Sci.* **2017**, *33*, 822–827. (In Chinese)
7. Sun, W.; Zhang, Q.; Qi, Y.N.; Dou, H. Condition Analysis on NDVI-Based Soybean Spatial Distribution and Growth in Argentina 2017. *China Sci. Technol. Resour. Rev.* **2017**, *49*, 66–75. (In Chinese)
8. Wei, P.F.; Xu, X.G.; Yang, G.J.; Li, Z.Y.; Wang, J.W.; Chen, G. Remote Sensing Classification of Crops Based on the Change Characteristics of Multi-phase Vegetation Index. *J. Agric. Sci. Technol.* **2019**, *21*, 54–61. (In Chinese)
9. Huang, Y.; Qiu, B.; Chen, C.; Zhu, X.; Wu, W.; Jiang, F.; Lin, D.; Peng, Y. Automated soybean mapping based on canopy water content and chlorophyll content using Sentinel-2 images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *109*, 102801. [CrossRef]
10. Chen, H.; Li, H.; Liu, Z.; Zhang, C.; Zhang, S.; Atkinson, P.M. A novel Greenness and Water Content Composite Index (GWCCI) for soybean mapping from single remotely sensed multispectral images. *Remote Sens. Environ.* **2023**, *295*, 113679. [CrossRef]
11. Zhou, J.P.; Li, C.J.; Shi, L.G.; Shi, S.; Hu, H.T.; Huai, H.J. Crops Distribution Information Extracted by Remote Sensing Based on Decision Tree and Object-oriented Method. *Trans. Chin. Soc. Agric. Mach.* **2016**, *47*, 318–326. (In Chinese)
12. Tian, F.Y.; Wu, B.F.; Zeng, H.W.; He, Z.X.; Zhang, M.; José, B. Identifying Soybean Cropped Area with Sentinel-2 Data and Multi-Layer Neural Network. *J. Geo-Inf. Sci.* **2019**, *21*, 918–927. (In Chinese)
13. De Souza, C.H.W.; Mercante, E.; Johann, J.A.; Lamparelli, R.A.C.; Uribe-Opa, M.A. Mapping and discrimination of soya bean and corn crops using spectro-temporal profiles of vegetation indices. *Int. J. Remote Sens.* **2015**, *36*, 1809–1824. [CrossRef]

14. Grzegozewski, D.M.; Johann, J.A.; Uribe-Opazo, M.A.; Mercante, E.; Coutinho, A.C. Mapping soya bean and corn crops in the State of Paraná, Brazil, using EVI images from the MODIS sensor. *Int. J. Remote Sens.* **2016**, *37*, 1257–1275. [CrossRef]
15. Wang, L.M.; Liu, J.; Yang, L.B.; Yang, F.G.; Fu, C.H. Application of Random Forest Method in Maize-soybean Accurate Identification. *Acta Agron. Sin.* **2018**, *44*, 569–580. (In Chinese) [CrossRef]
16. Liu, H.; Wu, W.B.; Shen, G.; Huang, Q. Spatio-temporal evolution of traditional soybean planting structure in Songnen Plain, China in 1996–2016. *Chin. J. Appl. Ecol.* **2018**, *29*, 3275–3282. (In Chinese)
17. Huang, J.X.; Hou, Y.Z.; Su, W.; Liu, J.M.; Zhu, D.H. Mapping corn and soybean cropped area with GF-1 WFV data. *Trans. Chin. Soc. Agric. Eng.* **2017**, *33*, 164–170. (In Chinese)
18. You, N.S.; Dong, J.W. Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 109–123. [CrossRef]
19. Liang, J.; Zheng, Z.W.; Xia, S.T.; Zhang, X.T.; Tang, Y.Y. Crop recognition and evaluationusing red edge features of GF-6 satellite. *Natl. Remote Sens. Bull.* **2020**, *24*, 1168–1179. (In Chinese) [CrossRef]
20. Ren, T.; Xu, H.; Cai, X.; Yu, S.; Qi, J. Smallholder Crop Type Mapping and Rotation Monitoring in Mountainous Areas with Sentinel-1/2 Imagery. *Remote Sens.* **2022**, *14*, 566. [CrossRef]
21. Shen, Y.; Li, Q.Z.; Du, X.; Wang, H. Indicative features for identifying corn and soybean using remote sensing imagery at middle and later growth season. *Natl. Remote Sens. Bull.* **2022**, *26*, 1410–1422. (In Chinese) [CrossRef]
22. Li, X.Y.; Yu, L.; Peng, D.L.; Gong, P. A large-scale, long time-series (1984–2020) of soybean mapping with phenological features: Heilongjiang Province as a test case. *Int. J. Remote Sens.* **2021**, *42*, 7332–7356. [CrossRef]
23. You, N.S.; Dong, J.W.; Huang, J.X.; Du, G.; Zhang, G.; He, Y.; Yang, T.; Di, Y.; Xiao, X. The 10-m crop type maps in Northeast China during 2017–2019. *Sci. Data* **2019**, *8*, 41. [CrossRef]
24. Liu, X.X.; Yu, L.; Zhong, L.H.; Hao, P.Y.; Wu, B.; Wang, H.S.; Yu, C.Q.; Gong, P. Spatial-temporal patterns of features selected using random forests: A case study of corn and soybeans mapping in the US. *Int. J. Remote Sens.* **2019**, *40*, 269–283. [CrossRef]
25. Wang, S.; Feng, W.; Quan, Y.; Li, Q.; Dauphin, G.; Huang, W.; Li, J.; Xing, M. A heterogeneous double ensemble algorithm for soybean planting area extraction in Google Earth Engine. *Comput. Electron. Agric.* **2022**, *197*, 106955. [CrossRef]
26. Cai, Y.; Guan, K.; Peng, J.; Wang, S.; Seifert, C.; Wardlow, B.; Li, Z. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sens. Environ.* **2018**, *210*, 35–47. [CrossRef]
27. Chen, S.Y.; Liu, J. Evaluation of deep learning algorithm for crop identification based on GF-6 time series images. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 161–168. (In Chinese)
28. She, B.; Yang, Y.Y.; Zhao, Z.G.; Huang, L.S.; Liang, D.; Zhang, D.Y. Identification and mapping of soybean and maize crops based on Sentinel-2 data. *Int. J. Agric. Biol. Eng.* **2020**, *13*, 171–182. [CrossRef]
29. Zhu, M.Q.; She, B.; Huang, L.S.; Zhang, D.Y.; Xu, H.F.; Yang, X.Y. Identification of soybean based on Sentinel-1/2 SAR and MSI imagery under a complex planting structure. *Ecol. Inform.* **2022**, *72*, 101825. [CrossRef]
30. Mcfeeters, S.K. The use of the Normalized Difference Water Index (*NDWI*) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
31. Zha, Y.; Gao, J.; Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Remote Sens.* **2003**, *24*, 583–594. [CrossRef]
32. Gong, P.; Liu, H.; Zhang, M.; Li, C.; Wang, J.; Huang, H.; Clinton, N.; Ji, L.; Li, W.; Bai, Y. Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. [CrossRef]
33. Zhang, D.Y.; Yang, Y.Y.; Huang, L.S.; Yang, Q.; Liang, D.; She, B.; Hong, Q.; Jiang, F. Extraction of soybean planting areas combining Sentinel-2 images and optimized feature model. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 110–119. (In Chinese)
34. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
35. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
36. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In *Goddard Space Flight Center 3d ERTS-1 Symposium*; NASA: Washington, DC, USA, 1974; Volume 1, Section A; p. 309.
37. Gitelson, A.A.; Merzlyak, M.N. Remote estimation of chlorophyll content in higher plant leaves. *Int. J. Remote Sens.* **1997**, *18*, 2691–2697. [CrossRef]
38. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
39. Gitelson, A.; Merzlyak, M.N. Spectral Reflectance Changes Associated with Autumn Senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. *J. Plant Physiol.* **1994**, *143*, 286–292. [CrossRef]
40. Sims, D.A.; Gamon, J.A. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sens. Environ.* **2002**, *81*, 337–354. [CrossRef]
41. Gitelson, A.A.; Gritz, Y.; Merzlyak, M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* **2003**, *160*, 271–282. [CrossRef]

42. Gitelson, A.A.; Keydan, G.P.; Merzlyak, M.N. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. *Geophys. Res. Lett.* **2006**, *33*, L11402. [CrossRef]
43. Gitelson, A.A. Wide Dynamic Range Vegetation Index for Remote Quantification of Biophysical Characteristics of Vegetation. *J. Plant Physiol.* **2004**, *161*, 165–173. [CrossRef] [PubMed]
44. Dwivedi, R.S.; Rao, B.R.M. The selection of the best possible Landsat TM band combination for delineating salt-affected soils. *Int. J. Remote Sens.* **1992**, *13*, 2051–2058. [CrossRef]
45. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
46. Immitzer, M.; Atzberger, C.; Koukal, T. Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data. *Remote Sens.* **2012**, *4*, 2661–2693. [CrossRef]
47. Wang, X.; Wang, Z.Q.; Jin, G.; Yang, J. Land reserve prediction using different kernel based support vector regression. *Trans. Chin. Soc. Agric. Eng.* **2014**, *30*, 204–211. (In Chinese)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.