Improved Accuracy of Phenological Detection in Rice Breeding by Using Ensemble Models of Machine Learning Based on UAV-RGB Imagery

Haixiao Ge 1,2, Fei Ma 1, Zhenwang Li 1, Zhengzheng Tan 3 and Changwen Du 1,2,*

1 The State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science Chinese Academy of Sciences, Nanjing 210008, China; gehaixiao@issas.ac.cn (H.G.); fma@issas.ac.cn (F.M.); zwli@issas.ac.cn (Z.L.)
2 College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
3 Yuan Longping High-Tech Agriculture Co., Ltd., Changsha 410001, China; tanzz@lpht.com.cn
* Correspondence: chwdu@issas.ac.cn; Tel.: +86-25-86881565; Fax: +86-25-86881000

Abstract: Accurate and timely detection of phenology at plot scale in rice breeding trails is crucial for understanding the heterogeneity of varieties and guiding field management. Traditionally, remote sensing studies of phenology detection have heavily relied on the time-series vegetation index (VI) data. However, the methodology based on time-series VI data was often limited by the temporal resolution. In this study, three types of ensemble models including hard voting (majority voting), soft voting (weighted majority voting) and model stacking, were proposed to identify the principal phenological stages of rice based on unmanned aerial vehicle (UAV) RGB imagery. These ensemble models combined RGB-VIs, color space (e.g., RGB and HSV) and textures derived from UAV-RGB imagery, and five machine learning algorithms (random forest; k-nearest neighbors; Gaussian naïve Bayes; support vector machine and logistic regression) as base models to estimate phenological stages in rice breeding. The phenological estimation models were trained on the dataset of late-maturity cultivars and tested independently on the dataset of early-medium-maturity cultivars. The results indicated that all ensemble models outperform individual machine learning models in all datasets. The soft voting strategy provided the best performance for identifying phenology with the overall accuracy of 90% and 93%, and the mean F1-scores of 0.79 and 0.81, respectively, in calibration and validation datasets, which meant that the overall accuracy and mean F1-scores improved by 5% and 7%, respectively, in comparison with those of the best individual model (GNB), tested in this study. Therefore, the ensemble models demonstrated great potential in improving the accuracy of phenology detection in rice breeding.

Keywords: UAV; machine learning; ensemble models; phenology; breeding

1. Introduction

In rice breeding, the accurate phenological information of rice is essential for breeders to make decisions in fertigation and breeding operations accordingly [1]. Crop phenological stages have a close relationship with climate conditions [2], but are also affected by genetic variability [3]. Traditionally, the approach of detecting crop phenology uses visual observations to identify the timing of periodic events such as germination, flowering, and physiological maturity [1,4]. However, field observation is indeed time-consuming and labor-intensive work which is often constrained by the observer’s subjective perception of the canopy status [5].

As an alternative method, remote sensing data based on satellites has been successfully used to detect landscape-scale phenology [2,6–8]. Landsat was firstly used to characterize the seasonal changes of vegetation at regional scales [9]. However, some plant phenological stages changed rapidly, which limited the application of Landsat due to its long revisit time (16 days). Since the Moderate Resolution Imaging Spectrometer (MODIS) was launched, it
Remote Sens. 2021, 13, 2678

provided time-series vegetation indices (VIs) with a high temporal (daily) and medium spatial resolution (250–500 m) [7]. Many previous studies have exploited the MODIS data to detect the phenological variation of different crops including maize [7], wheat [10,11] and rice [12–14], as well as other crops [15–17]. In addition, some studies retrieved plant phenology from fine resolution satellite imagery based on new optical sensors [18]. Pan et al. [6] constructed a time-series NDVI dataset from HJ-1 A/B and found that the crop season start or end from HJ-1 A/B imagery was comparable with local observations. Vrieling et al. [18] mapped the spatial distribution of vegetation phenology at 10 m resolution derived from Sentinel-2A. However, the application of satellite-optical sensors has still been challenged as a method of phenology retrieval due to the risk of poor data quality as a result of clouds [19]. Synthetic Aperture Radar (SAR) satellites without the limitation of weather conditions have been applied to derive vegetation morphological and phenological information [20]. Fikriyah et al. [21] used temporal information based on SAR backscatter to monitor the differences in rice growth. He et al. [22] found that the backscattering parameters from RADARSAT-2 had a strong correlation with rice phenology in a study in Meishan, China. Although the regional or global distributions of vegetation phenology can be retrieved from satellite sensors, these are not capable of properly detecting phenological variability of small-size farmland and crop breeding due to the limitation of high temporal and spatial resolution [1,4,23].

Compared to the satellites, the imagery from digital cameras for near-surface observations had a higher temporal and spatial resolution [24]. Especially, the consumer-grade cameras with RGB (R, G, B) bands mounted on unmanned aerial vehicles (UAV) are cost-effective and flexible to revisit fields to collect time-series images during the vegetation growing season [4]. Park et al. [25] used RGB images from UAV with a machine learning method to assess leaf phenology of trees and highlighted the feasibility of texture characteristics for phenological analysis in tropical forest. Burkart et al. [26] calculated the time-series Green-Red Vegetation Index (GRVI) from UAV-RGB imagery and analyzed its potential to track seasonal changes of barley. Yang et al. [4] constructed a deep learning approach to identify principal phenology of rice from a mono-temporal UAV imagery. These preliminary studies have proven the feasibility of tracking vegetation phenology based on UAV imagery with various working principles.

By reviewing the previous studies in the literature, the common methodologies of phenology acquisition from satellite or near-surface images are mainly shown as three types: (1) curve fitting methods which used logistic function, Gaussian function and Savitzky–Golay filtering to smooth the VI’s time-series curves and then extracted phenology based on its inflection points [27–29]; (2) shape model fitting (SMF) methods which smoothed VI data and extracted phenological stages based on optimum scaling parameters and visual observations [7,24]; (3) machine learning or deep learning algorithms which fed VIs, textures or images into models to estimate phenological stages [4,25]. The first two types of methods highly relied on the amount of time-series image data and neglected the heterogeneity of crops’ annual growth progress [4,24]. Since the advance of computing power and neural network algorithms, many studies have successfully used deep learning techniques to estimate vegetation parameters [30–33]. However, the large amount of reference data for deep learning is still a challenge for scientists. In contrast, machine learning has no high requirements for computer hardware configuration and amount of datasets. Therefore, machine learning has great potential to reconstruct time-series datasets and detect phenology [24]. In recent years, ensemble machine learning models based on a combination of individual models have been proved to be more effective at making predictions [34,35]. Shahhosseini et al. [36] combined linear, least absolute shrinkage and selection operator (LASSO), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and random forests (RF) to forecast corn yield. Masiza et al. [37] combined XGBoost with support vector machine (SVM), artificial neural network (ANN), naive Bayes (NB) and RF to map landscape in a smallholder farming system. Although some ensemble methods have been applied in agriculture domain, to the best
of our knowledge, the performance and potential of different ensemble models are rarely compared and tested for detecting phenology.

In this study, our main research problem is as follows: is UAV-RGB imagery sufficient to accurately identify phenology in heterogeneous rice? Moreover, given the recent introduction of ensemble algorithms into phenological extraction, little is known about whether ensemble models could improve the accuracy in phenological detection compared to machine learning models. Based on the above-mentioned question, firstly, the high-resolution imagery from UAV-RGB images was used to calculate VIs and textures in the plot level. Then, three model ensemble methods (soft voting, hard voting and model stacking) based on five machine learning algorithms (k-nearest neighbors, Gaussian naïve Bayes, logistic regression, RF and SVM) were the first time to be used to detect rice phenology in breeding trials. Accordingly, the objectives of the study are: (1) to build the ensemble models to detect rice phenology, (2) to compare the performance of the machine learning models and ensemble models for detecting phenology, and (3) to analyze the contributions of feature variables in the ensemble models.

2. Materials and Methods

2.1. Study Area

The breeding trails were carried out at the Guanshan rice breeding base (28°19’ N, 112°40’ E), Jinzhou town, Ningxiang district, Changsha city, Hunan Province, China (Figure 1). The annual average rainfall was 1358 mm. Average daily temperature was 16.8 °C. The average daily sunshine hour duration is about 4.8 h. There are two hybrid rice experiments in the study area. Experiment 1 was designed with late-maturity cultivars and sowing date was 18 June 2020. Experiment 2 was designed with early-medium-maturity cultivars which were sowed at 22 June 2020. The experimental field was divided into 437 plots in total. Each cultivar was transplanted into an individual plot. Among these plots, 279 plots and 158 plots were the same sizes of 2.5 × 5 m and 2.5 × 7.5 m, respectively. The same N and irrigation treatment was applied for all plots. Eleven ground control points (GCPs) were distributed across the field (Figure 1). These points were obtained from a Hiper Topcon DGPS (Topcon Co., Ltd., Tokyo, Japan) with 0.01 m vertical and horizontal precision.

2.2. Field-Based Observations of Rice Phenology

In the field survey, rice breeders observed and recorded the phenological stages of the different cultivars during rice growing season. In this study, the vegetative phase, initial heading stage, heading stage, full heading stage, ripening phase and maturity were collected (Figure 2). The detailed descriptions of these phenological stages were presented by Fageria [38] and Yang et al. [4].
2.3. Image Preprocessing

The UAV used in this paper was a four-rotor consumer-grade UAV called Phantom 4 Pro (DJI Technology Co., Ltd., Shenzhen, China) which was equipped with a RGB camera offering a spatial resolution of 20-mega pixels. A lens of 35 mm focal length was mounted on the RGB camera which could acquire an image with \(5456 \times 3632\) pixels in a short period of time (faster than \(1/2000\) s). The flight duration of the UAV was about 30 min. During rice growing season, five flights were carried out between August and November in 2020. The detailed information of flights is shown in Table 1. The UAV was flown at the flight
altitude of 50 m above ground for each flight, resulting in the ground sampling distance of 1.36 cm. At each flight, Pix4Dcapture software was used to set the flight waypoints. This resulted in the flight speed of 2.8 m/s, forward overlap of 80% and side overlap of 70%. In addition, the flights were conducted on the stable ambient light between 11:00 am and 13:00 pm [39].

Table 1. Dates of the UAV flights and statistical numbers of plots in main phenological stages for each flight.

<table>
<thead>
<tr>
<th>Date of UAV Flights</th>
<th>Vegetative</th>
<th>Initial Heading</th>
<th>Heading</th>
<th>Full Heading</th>
<th>Ripening</th>
<th>Maturity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 August 2020</td>
<td>437</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>29 August 2020</td>
<td>417</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>11 September 2020</td>
<td>5</td>
<td>78</td>
<td>148</td>
<td>91</td>
<td>115</td>
<td>–</td>
</tr>
<tr>
<td>30 September 2020</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>437</td>
<td>–</td>
</tr>
<tr>
<td>1 November 2020</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>75</td>
<td>362</td>
</tr>
</tbody>
</table>

2.3.2. Generation of Orthomosaic Maps

After each flight, the RGB images of the experimental plots were downloaded from the SD card. Generating orthomosaic maps requires one to (1) align raw images; (2) georeference images—11 GCPs were imported, placed manually on one image and automatically projected to the remaining images [40]; and (3) export the orthomosaic maps with TIFF-image format. In this study, Pix4Dmapper 1.1.38 (Pix4D SA, Lausanne, Switzerland) was used to carry out the above-mentioned steps to generate orthomosaic maps. The radiance reaching the camera had an empirical linear relationship with digital numbers (DN) in each band; therefore, linear equations which were introduced by Yu et al. [41] can be carried out to maintain radiometric consistency of multi-temporal images. The images taken on August 1 were used as the reference images; normalization coefficients for images taken at subsequent dates were calculated and applied to the orthomosaic maps to accomplish radiometric normalization.

2.3.3. Calculation of VIs or Color Space

The VIs or color space were calculated from the DN values of the UAV-RGB imagery (see Table 2). Firstly, ArcGIS v.10.0 (ESRI, Redlands, CA, USA) was used to generate regions of interest (ROIs) by manually delineating plots in orthomosaics (Figure 1). Secondly, the selected VIs or color space were extracted based on the R, G, B bands of orthomosaic maps by using Python 3.6. Then, the mean VIs or bands of ROI were calculated by using the “ZonalStatisticsAsTable” module in ArcGIS v.10.0 software.

Table 2. Descriptions of VIs or color space derived from UAV-RGB imagery.

<table>
<thead>
<tr>
<th>Type</th>
<th>Variable</th>
<th>Formula/Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIs</td>
<td>Normalized red index (NRI)</td>
<td>R/(R + G + B)</td>
<td>[42]</td>
</tr>
<tr>
<td></td>
<td>Normalized green index (NGI)</td>
<td>G/(R + G + B)</td>
<td>[42]</td>
</tr>
<tr>
<td></td>
<td>Normalized blue index (NBI)</td>
<td>B/(R + G + B)</td>
<td>[42]</td>
</tr>
<tr>
<td></td>
<td>Normalized excess green index (ExG)</td>
<td>(2G − R − B)/(G + R + B)</td>
<td>[43]</td>
</tr>
<tr>
<td></td>
<td>Normalized excess red index (ExR)</td>
<td>(1.4R − G)/(G + R + B)</td>
<td>[44]</td>
</tr>
<tr>
<td></td>
<td>Green and red index (VIgreen)</td>
<td>(G − R)/(G+R)</td>
<td>[45]</td>
</tr>
<tr>
<td>Color space</td>
<td>R band of UAV image (R)</td>
<td>DN values of R band</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>G band of UAV image (G)</td>
<td>DN values of G band</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>B band of UAV image (B)</td>
<td>DN values of B band</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Hue (H)</td>
<td>The DN values of hue</td>
<td>[46]</td>
</tr>
<tr>
<td></td>
<td>Saturation (S)</td>
<td>The DN values of saturation</td>
<td>[46]</td>
</tr>
<tr>
<td></td>
<td>Value (V)</td>
<td>The DN values of value</td>
<td>[46]</td>
</tr>
</tbody>
</table>
2.3.4. Calculation of Texture Features

The methodology of gray level co-occurrence matrix (GLCM) was proposed by Haralick et al. [47] for characterizing the heterogeneity and inhomogeneity of images. GLCM highlights texture properties, such as homogeneity, contrast or irregularity. Texture was the term which was used to characterize the tones of gray-level variations in digital images [48]. The construction of GLCM firstly converted an image to grayscale to be discretized in the whole matrix. The method was to divide the range of continuous pixel values into N sub samples of equal width, called gray level, and then remap these values on a single gray level. Based on the discretized mapping, the elements of GLCM were obtained by calculating the frequency of pixel pairs with specific gray levels and spatial relationship in the matrix. Many previous studies have successfully extracted GLCM from UAV-based imagery to estimate crop parameters [49–51]. In this study, ENVI 5.3 (Harris Geospatial Solutions, Inc, Broomfield, CO, USA) was applied to calculate eight textures (mean, variance, contrast, homogeneity, dissimilarity, correlation, entropy and second moment). To calculate the texture for each pixel, a $3 \times 3$ window size as a moving window was used to obtain statistics. This selected window size could capture small differences between pixel values in small extents [52]. After calculation of GLCM, the ArcGIS v.10.0 software was used to obtain the mean of each texture for each plot.

2.4. Classifier Techniques
2.4.1. Multiple Machine Learning Algorithms

Five machine learning classification algorithms were selected to detect phenology in rice breeding. These algorithms are including RF, k-nearest neighbors (KNN), Gaussian naïve Bayes (GNB), support vector machine (SVM) and logistic regression (LR). All the user-defined parameters from the five different types of classifiers were determined by “GridSearchCV” method with 10-fold cross validation in Python 3.6 environment (Table A1). Brief descriptions of the selected algorithms are in the following paragraphs.

RF is a widely used machine learning algorithm which was developed by Breiman [53]. RF was created based on the combination with a set of decision trees. The classification results of RF were determined by majority voting of trees. In this study, RF has two parameters to be tuned including “max_depth” (the maximum depth of trees) and “n_estimators” (the number of trees in the forest). The “RandomForestClassifier” function in “sklearn” package was used to implement RF.

KNN is a well-known machine learning method for classification [54]. The basic idea of this algorithm is that most of the K similar samples (neighbors) in the feature space belong to a certain category, and then the sample belongs to this category. The classifiers of samples were determined by the category of the nearest one or several samples. For this study, only one parameter defined by the user, “n_neighbors” (number of neighbors) needs to be tuned in “sklearn” package.

GNB was one of simple fast classification algorithms, which was usually suitable for datasets with high dimensions [55]. Because of its fast running speed and few tuning parameters, it is very suitable to provide fast, robust and automated results for classification [55]. In this paper, default parameters were used to train the GNB model.

SVM used a kernel learning method to classify data, even if the data are not linearly separable [56]. For the nonlinear problems, the method of SVM is to select the kernel function and map the training samples to the high-dimensional space through the kernel function, so as to obtain the best hyperplane in the high-dimensional space. The advantage of using the kernel function is that the computation is not much more complex than before while obtaining the optimal hyperplane [54,57]. Compared with other traditional learning algorithms, the SVM algorithm can effectively avoid the problems of local optimal solution and dimension disaster. In this work, two user-defined parameters including “C” (the strength of the regularization) and “kernel” (the kernel type) were tuned.

LR is a widely accepted algorithm which can deal with binary or multiclass problems [58]. In this study, the “LogisticRegression” function in the “sklearn” package was
used to obtain the LR model. The LR algorithm was applied with three user-defined parameters including “C” (inverse of regularization strength), “penalty” (specify the norm used in the penalization) and “solver” (the class implements regularized logistic regression).

2.4.2. Ensemble Models

The primary goal of using ensemble models is to reduce the risk of selecting a classifier with poor performances, or improve the accuracy of an individual classifier [59]. In this study, three ensemble strategies, namely hard voting, soft voting and model stacking were used to create ensemble models (Figure 3). In the case of hard voting, the output of the ensemble model is the most assigned class by multiple classifiers. In the soft voting rule, a weight was assigned to individual classifier to favor classifiers with better performance in voting decision. Model stacking takes predictions generated by multiple classifiers and uses these predictions as inputs in a second-level learning classifier.

![Figure 3. Schematic diagram of the proposed methods.](image)

The hard voting rule of an ensemble model on decision problem \( x \) works as follows:

\[
D_x = \max_{idx} \sum_{i=1}^{L} d_{x,i}
\]

(1)

where \( D_x \) is the selected class number of hard voting on decision problem \( x \). \( L \) is the number of voters (base classifiers). \( d_{x,i} \) is the binary vector output (e.g., \([0, \ldots, 0, 1, 0, \ldots, 0]\)) of the classifier on the decision problem \( x \). \( \max_{idx} \) is the index where the maximum value was obtained in the summation vector \( \sum d_{x,i} \). If multiple maximum values were found, the index of first occurrence is returned.

For the soft voting, the classifiers with good performances were assigned larger weights. The soft voting is an extension of the Equation (1) and uses weights \( \omega_i \) per base machine learning classifier \( i \).

\[
D_x = \max_{idx} \sum_{i=1}^{L} w_i d_{x,i}
\]

(2)

\[
w_i = \log[k/(1-k)]
\]

(3)

where \( k \) is the kappa coefficient of the classifier \( i \) over the independent sample set [60].
Model stacking is an ensemble algorithm that assembles RF, KNN, GNB, SVM and LR methods. According to this method, a higher level (level 2) meta-model is trained on the predictions of base models (level 1); then, a final output of the overall ensemble model is generated [61]. In this study, five machine learning models were used as the level-1 models and GNB as the level-2 model.

2.5. Evaluation of Model Accuracy

In order to evaluate the performance of different machine learning classifiers and ensemble models for phenology detection, several strategies were carried out. Firstly, the accuracy of each machine learning algorithm was tested. Secondly, the proposed ensemble models based on five individual machine learning classifiers were conducted for detecting rice phenology. Then, the ensemble models were compared with the result of the best individual classifier model. Finally, the phenological maps in rice breeding produced by the optimal model were visually evaluated.

In this study, the data from Experiment 1 were selected to train the estimation models, and the independent data from Experiment 2 were used to validate these models. The assessment of models was carried out on the six principal phenological stages: vegetative stage, initial heading stage, heading stage, full heading stage, ripening stage and maturity stage. Two metrics were selected to evaluate the accuracies of classifiers: overall accuracy and F1-score. These two metrics are calculated as follows:

\[
\text{overall accuracy} = \frac{(TP + TN)}{N} \tag{4}
\]

\[
\text{Recall} = \frac{TP}{(TP + FN)} \tag{5}
\]

\[
\text{Precision} = \frac{TP}{(TP + FP)} \tag{6}
\]

\[
F1\text{-score} = \frac{(2 \times \text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{7}
\]

where \(TP, TN, FP, FN\) and \(N\) in Equation (4) represent True Positive, True Negative, False Positive, False Negative and a total number of items, respectively. The confusion matrixes were used to show the distribution of omission and commission errors across different phenological stages. In addition, the “feature_importance_permutation” function in “mlxtend” package was used to evaluate the feature importance of ensemble models in Python 3.6 running environment.

3. Results

3.1. Variation of Observed Phenology in Different Plots

Figure 4 shows the statistical distributions of observed main phenology in different plots. Although there was a four-day difference in sowing dates between Experiment 1 and Experiment 2, all plots had the same transplanting date (13 July 2020). Due to the variance of cultivar traits, the phenology was highly heterogeneous in different plots. The dates of phenology in early-medium-maturity cultivars were earlier than those in late-maturity cultivars. The mean dates of initial heading, full heading and maturity stages in Experiment 1 are September 8 (DOY = 252), September 14 (DOY = 258) and November 11 (DOY = 306), respectively. The mean dates of corresponding phenology in Experiment 2 are September 4 (DOY = 248), September 10 (DOY = 254) and October 26 (DOY = 300), respectively.

3.2. Model Performance by Multiple Machine Learning

Five machine learning algorithms, including RF, KNN, GNB, SVM and LR, were investigated to compare the performance for phenological detection. In the calibration dataset, RF performed best among these machine learning models (overall accuracy = 93% and mean F1-score = 0.84). Other models produced comparable accuracies, resulting in overall accuracies from 85% to 88% and mean F1-scores from 0.73 to 0.77 (Figure 5). For individual phenology, all models had good performance for detecting vegetative, ripening and maturity stages and resulted in F1-scores from 0.81 to 1.0 (Table 3). However, these
models obtained slightly worse accuracies for phenological detection at initial heading, heading and full heading stages. Accordingly, the F1-scores ranged from 0.39 to 0.79.

Figure 4. The statistical distribution of the main observed phenology in Experiment 1 (late-maturity cultivars) and Experiment 2 (early-medium-maturity cultivars).

In the validation dataset, the GNB model obtained the most accurate estimation with overall accuracy of 88% and mean F1-score of 0.74. The RF, LR and SVM models had similar performance with overall accuracies ranging from 85% to 87%, and mean F1-scores ranging from 0.72 to 0.73. The KNN model yielded the worst performance (overall accuracy = 81% and mean F1-score = 0.57). Except for the KNN model, other models tested in the validation dataset showed consistency with those in the calibration dataset for detecting vegetative, ripening and maturity stages (F1-scores from 0.83 to 0.98). The transition dates of initial heading, heading and full heading are close to each other. It is challenging for all models to accurately identify these phenological stages. Thus, the F1-scores of different machine
learning models at heading phase were lower than those at other phenological stages and ranged from 0.44 to 0.65 except for KNN.

Table 3. The F1-scores of different machine learning models for detecting phenology in calibration and validation datasets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Models</th>
<th>Phenological Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Vegetative</td>
</tr>
<tr>
<td>Calibration</td>
<td>RF</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>1.00</td>
</tr>
<tr>
<td>Validation</td>
<td>RF</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.97</td>
</tr>
</tbody>
</table>

3.3. Model Performance by Ensemble Models

Three different strategies of ensemble models based on base classifiers (five machine learning models) were investigated to evaluate the performances for phenological detection in rice breeding. Figure 6 shows the modeling performance of ensemble models (hard voting, soft voting and model stacking) for monitoring phenology in calibration and validation datasets. In the calibration dataset, the overall accuracies and mean F1-scores of ensemble models ranged from 90% to 92% and 0.79 to 0.82, respectively. All the ensemble models had more accurate performances for detecting vegetative, ripening, and mature stages (F1-scores ranging from 0.86 to 1.0) than those for heading phase (F1-scores ranging from 0.54 to 0.78) (Table 4).

![Figure 6](image-url)
Table 4. The F1-scores in different ensemble models for detecting phenology in calibration and validation datasets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Models</th>
<th>Phenological Stage</th>
<th>Vegetative</th>
<th>Initial Heading</th>
<th>Heading</th>
<th>Full Heading</th>
<th>Ripening</th>
<th>Maturity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration</td>
<td>Hard voting</td>
<td>1.00</td>
<td>0.78</td>
<td>0.72</td>
<td>0.92</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soft voting</td>
<td>1.00</td>
<td>0.54</td>
<td>0.75</td>
<td>0.66</td>
<td>0.90</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model stacking</td>
<td>1.00</td>
<td>0.64</td>
<td>0.78</td>
<td>0.59</td>
<td>0.94</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>Hard voting</td>
<td>0.97</td>
<td>0.67</td>
<td>0.74</td>
<td>0.92</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soft voting</td>
<td>0.97</td>
<td>0.53</td>
<td>0.65</td>
<td>0.80</td>
<td>0.96</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model stacking</td>
<td>0.96</td>
<td>0.42</td>
<td>0.60</td>
<td>0.62</td>
<td>0.91</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>

In the validation dataset, the soft voting model (overall accuracy = 93% and mean F1-score = 0.81) outperformed the hard voting (overall accuracy = 92% and mean F1-score = 0.79) and model stacking (overall accuracy = 89% and mean F1-score = 0.74). As for individual phenology, the soft voting was slight better than the hard voting in terms of F1-scores, especially at the late stage of rice growth (Table 4). In addition, the model stacking strategy obtained worse accuracies than those in the hard voting and soft voting models for detecting the same phenological stage.

3.4. Single Best Model vs. Ensemble Models

Among five machine learning models, the GNB model with the best performance in terms of overall accuracy and mean F1-score in the validation dataset was selected to compare the reliability of phenological detection with ensemble models. Figure 7 presents the classification results obtained by GNB and ensemble models. It was observed that the vegetative stage was the most easily identified among the six phenological stages for all models. The heading stage in ensemble models was less confused with full heading stage than in the GNB model. Ripening phase is less confused with full heading stage as well. However, the capability for detecting full heading and maturity stage in the GNB model was slightly better than those in ensemble models.

Figure 8 shows the differences in F1-scores between ensemble models and the GNB model. In the calibration dataset, all the ensemble models showed improvements of 2–15% for individual phenology detection in F1-scores relative to the performance of the GNB model. The ensemble models increased overall accuracies and mean F1-scores up to 6–9% and 5–7%, respectively, compared to the GNB model. In the validation dataset, all ensemble models had somewhat poor performance for maturity detection with the F1-scores reduced by 2.2–5.1%. As for overall accuracy and mean F1-score, the proposed ensemble models offered an improved accuracy in classification results with overall accuracy increased by 1–5%, and mean F1-scores increased by 0.2–7%.
Figure 7. The confusion matrices of simulated and observed phenology achieved by GNB (a), hard voting (b), soft voting (c) and model stacking (d) using the validation dataset. Note: 0–5 represent vegetative, initial heading, heading, full heading, ripening and maturity stages, respectively.

In general, all ensemble models improved the classification accuracy to a certain extent, compared to the single best model (GNB). In addition, the highest classification accuracy was acquired by the soft voting algorithm, followed by the hard voting algorithm and finally by the model stacking.
Figure 8. Differences in F1-scores (%) between ensemble models and the single best classifier (GNB) in calibration (a) and validation (b) datasets.

4. Discussion

4.1. Feasibility of Ensemble Models for Phenology Detection

Many previous studies have successfully used the time-series VI data to capture vegetation phenology based on the function fitting or threshold method [6,18,61]. However, the accuracy of monitoring phenology is largely constrained by the temporal resolution of images [4]. In contrast, using machine learning algorithms or ensemble models to carry out phenology detection is another strategy without the limitation of temporal resolution, as long as there are enough training samples to identify each phenological stage. In this study, the performance of five machine learning algorithms and three ensemble models was compared to detect phenology in rice breeding. In terms of overall accuracy and F1-scores, all ensemble models were more accurate than the five individual classifiers in both calibration and validation datasets. Our results emphasized that the strategies to combine individual classifiers had the potential to improve phenology classification. This is consistent with literature that finds that ensemble models outperformed individual machine learning algorithms in many applications, such as mapping smallholder fields [37], predicting invasive plants [62] and forecasting corn yield [36].

Among these ensemble models, the soft voting (weighted majority voting) performed best. In addition, maps produced by the soft voting appeared more homogeneous than the best individual classifier (GNB) in visual analysis (Figure 9). This is because the base classifiers which are more effective at detecting phenology would have more influence on the combined outputs. Polikar [59] also pointed out that the soft voting by weighting the decisions of qualified classifiers more heavily could improve the overall performance of classification more so than the hard voting (majority voting). The hard voting which obtained results based on the highest number of votes tended to be less accurate, most likely due to the strong correlations of predictions in base models. As for model stacking, GNB was only used as the meta-classifier, and the accuracy was slightly worse than the soft voting. The potential of using other machine learning or deep learning algorithms to fuse individual classifiers was necessarily investigated subsequently.
Although the combination of individual classifiers can provide better phenological classification than a single classifier, the similarity of morphological characteristics in heading phase is still a challenge for ensemble models to exactly identify initial heading and full heading stages. In addition, the imbalanced training data could lead to poor classification [61]. Accordingly, the more representative samples in heading phase need to be further collected to improve the accuracy of ensemble models.

4.2. Combination of Feature Variables

In this study, three types (RGB-VIs, color space, textures) of feature variables were extracted from the UAV-RGB images to establish phenological prediction models. Although we did not compare the differences of performance of models based on RGB-VIs only, color space only and textures only, respectively; many previous studies have shown that fusion of multiple types of feature variables can improve the estimation accuracy of models [49,51,63,64]. Our research also confirms this point. The top five important variables come from different types of variables in the three ensemble models (Table 5). In addition, other auxiliary data derived from UAV-RGB imagery, such as 3D point cloud and canopy coverage of each plot, should be added as they may result in higher estimation accuracy. This is something which needs to be investigated in future with procedures to estimate rice phenological stages using more input features. Furthermore, the proposed
ensemble models could be used to obtain insight into critical factors which determine the variability of phenology in rice breeding for breeders and agronomists.

Table 5. Feature ranking in ensemble models based on variable importance.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Features</th>
<th>Soft Voting</th>
<th>Hard Voting</th>
<th>Model Stacking</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB-VIs</td>
<td>1</td>
<td>VIgreen</td>
<td>9</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>ExG</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>ExR</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>NRI</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>NGI</td>
<td>15</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>NBI</td>
<td>19</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>Color space</td>
<td>7</td>
<td>R</td>
<td>13</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>G</td>
<td>18</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>B</td>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>H</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>S</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>V</td>
<td>17</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>Texture</td>
<td>13</td>
<td>Contrast</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Correlation</td>
<td>20</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Dissimilarity</td>
<td>10</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>Entropy</td>
<td>14</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>Homogeneity</td>
<td>12</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>Mean</td>
<td>16</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>Second moment</td>
<td>11</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Variance</td>
<td>6</td>
<td>5</td>
<td>9</td>
</tr>
</tbody>
</table>

4.3. Potential of UAV-Based Imagery in Rice Breeding Trails

In recent years, the rapid development of UAVs and sensors along with improved computer capacities for data analyses has enabled a range of possibilities to estimate plant traits in high throughput field phenotyping [65]. Compared to fixed-position cameras and satellites, the UAV platforms had a better trade-off between revisit time, spatial resolution, price and data accuracy [66]. In particular, the phenotyping occurs at a relatively small plot scale in breeding trails. The flight altitudes of UAVs can be flexibly set by the requirements of experiments to capture ultra-high resolution images. The high spatial resolution of field phenotyping is a prerequisite to distinguish the changes of morphology in rice breeding which failed to be detected by satellites. In this study, the imagery with an ultra-high resolution of 1.36 cm was used to calculated feature variables. This resolution can track the temporal pattern of canopy color changes during crop growth season. In future, integrating the phenotypic information derived from UAVs into crop models to help decision making in agriculture management could bring large benefits in optimizing resource utilization to improve breeding efficiency.

5. Conclusions

In this study, three types of ensemble model were proposed to identify principal phenological stages in rice breeding trials using UAV-based RGB imagery. In contrast to the curve fitting or shape model fitting methods based on time-series VI data, the ensemble models could detect phenology from the mono-temporal UAV imagery. Results showed that all the ensemble models showed improvements of 1–5% in overall accuracy and 0.2–7% in mean F1-scores compared to the best individual classifier (GBN) for phenology detection in the validation dataset. Among the ensemble models, the soft voting had the best performance, followed by the hard voting and finally by the model stacking. The ensemble techniques presented in this study are promising and valuable to improve the accuracy of phenology detection in rice breeding, thus promoting the breeding processes.
**Author Contributions:** Conceptualization, H.G.; methodology, H.G.; investigation, C.D., H.G., F.M., Z.L. and Z.T.; resources, C.D.; data curation, H.G.; writing—original draft preparation, H.G.; writing—review and editing, C.D.; visualization, H.G.; supervision, C.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was supported by the Key Research and Development Program of Shandong Province (2019ZZY010713) and National Key Research and Development Program of China (No. 2018YFE0107000).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing not applicable.

**Acknowledgments:** Thanks to Jingyou Guo and Youdong Hu for providing the details of the experimental design in this study.

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

**Appendix A**

**Table A1.** Description of tuning parameters in the five machine learning algorithms with 10-fold cross validation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters Range</th>
<th>The Optimal Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>max_depth 10–100 (interval: 10)</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>n_estimators 3–15 (interval: 1)</td>
<td>6</td>
</tr>
<tr>
<td>KNN</td>
<td>n_neighbors 1–11</td>
<td>5</td>
</tr>
<tr>
<td>GNB</td>
<td>C –</td>
<td>default</td>
</tr>
<tr>
<td>SVM</td>
<td>kernel ‘linear’ and ‘rbf’</td>
<td>‘linear’</td>
</tr>
<tr>
<td></td>
<td>C 1–11</td>
<td>2</td>
</tr>
<tr>
<td>LR</td>
<td>penalty ‘L1’ and ‘L2’</td>
<td>‘L1’</td>
</tr>
</tbody>
</table>

**References**

1. Ma, Y.; Jiang, Q.; Wu, X.; Zhu, R.; Gong, Y.; Peng, Y.; Duan, B.; Fang, S. Monitoring hybrid rice phenology at initial heading stage based on low-altitude remote sensing data. *Remote Sens.* 2021, 13, 86. [CrossRef]


35. Hieu, P.; Olaffson, S. Bagged ensembles with tunable parameters. J. Classif. 2020, 37, 105398. [CrossRef]


57. Wozniak, M.; Grana, M.; Corchado, E. A survey of multiple classifier systems as hybrid systems. *Inf. Fusion* 2014, 16, 3–17. [CrossRef]
58. Wozniak, M.; Grana, M.; Corchado, E. A survey of multiple classifier systems as hybrid systems. *Inf. Fusion* 2014, 16, 3–17. [CrossRef]
59. Wozniak, M.; Grana, M.; Corchado, E. A survey of multiple classifier systems as hybrid systems. *Inf. Fusion* 2014, 16, 3–17. [CrossRef]