remote sensing

MDPI

Article

Machine Learning Fusion Multi-Source Data Features for Classification Prediction of Lunar Surface Geological Units

Wei Zuo [1,2], Xingguo Zeng [1], Xingye Gao [1], Zhoubin Zhang [1], Dawei Liu [1] and Chunlai Li [1,2,*]

[1] Key Laboratory of Lunar and Deep Space Exploration, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: licl@nao.cas.cn

Abstract: Taking the Chang'e-4 and Chang'e-5 landing areas as the study areas, this study extracts the geological unit information from the regional USGS geological map, as well as the feature information such as topography and geomorphology, material composition and mineral abundance from Chang'e-2 DOM and DEM, wide angle camera (WAC) and Kaguya multi-band imager data. By applying methods including the statistical-based estimation of mutual information of data and the integrated-algorithmic-model-based evaluation of feature importance to this extracted information, we screen the significant features and construct a high-precision classification model by combining machine learning algorithm with important features of sample data. The practical application of the multi-classification prediction on the complex geological units in the two study areas achieves 97.9% and 95.1% accuracy. At the same time, the significant characteristics of the study area are mined, and the rules and knowledge associated with the geological evolution of the study area are obtained. In this study, we carry out research on quantitative prediction and identification of lunar surface geological units based on large samples and construct a high-precision multi-classification model to achieve automatic classification and prediction on large sample geological units with high accuracy. This method provides a new idea for the predicted mapping of geological units of lunar global digital mapping. In addition, it helps to fully exploit the useful information in the data and enrich the knowledge regarding the formation and evolution of the Moon.

Keywords: data mining; lunar surface geological units; machine learning; multi-classification supervised learning; information fusion; the Moon

Citation: Zuo, W.; Zeng, X.; Gao, X.; Zhang, Z.; Liu, D.; Li, C. Machine Learning Fusion Multi-Source Data Features for Classification Prediction of Lunar Surface Geological Units. *Remote Sens.* **2022**, *14*, 5075. https://doi.org/10.3390/rs14205075

Academic Editor: Giancarlo Bellucci

Received: 15 September 2022
Accepted: 9 October 2022
Published: 11 October 2022

1. Introduction

The Moon, as the closest celestial body to the Earth, is the preferred target for human deep space exploration. Since the 1960s, the US Geological Survey (USGS) has produced a lunar geologic map using images obtained by the five Lunar Orbiter missions. Subsequently, a 1:5,000,000 global lunar geologic map was created combined with digital terrain model (DTM) data from the LRO, LOLA and Selene Kaguya missions, and it was publicly released in 2020 [1] for downloading (https://bit.ly/LunarGeology (accessed on 3 March 2020)). To date, it remains the most complete and latest global lunar geologic map publicly available. This map consists of 49 geologic units across the entire lunar surface. These units are broken down into groups based on attributes and include materials of craters, basins, terra, plains, Imbrium Formation, Orientale Formation and volcanic units [1]. A geological unit is a geological body with the same origin formed in a certain region by a defined geological activity in a specific period of time. The classification and delineation of lunar geological units is not only an essential and fundamental endeavor for carrying out lunar geological mapping, but also the basis for in-depth research on the integrity and regularity of the origin and evolution of the Moon [2,3].

The classification and delineation of lunar geological units implies the process of obtaining information on the spatial distribution, origin and evolution of lunar geological

units based on the summary of various lunar exploration data. The lunar surface geological unit is a comprehensive expression of the lunar surface's morphological features, material composition, mineral distribution, albedo, geological age, etc., reflecting the lunar formation and evolution processes [4,5]. With the rapid development of machine learning in recent years, research on lithology identification, lithological unit mapping and other related classification problems based on machine learning have achieved better results and progress [6–16]. Compared to traditional geological mapping techniques, the classification models or combination algorithms of machine learning are efficient and intelligent in lithology classification and recognition, and they can be used as an auxiliary tool with great potential advantages to improve the efficiency of the traditional geological mapping technology system. In the field of lunar geological mapping, the application of machine learning methods is still in the initial stage, and the classification and delineation of geological units are mainly focused on the methods combining the traditional GIS technology with the lunar exploration data. The application of machine learning methods in the field of lunar and planetary mapping is in the ascendant. The research on geological unit classification based on machine learning is not only an exploration of the method of lunar geological unit mapping, but also a way to extract the association rules from the data to obtain new knowledge and make discoveries, as well as to enrich and deepen the cognition of the formation and evolution of the Moon.

In this work, we extract information including lunar surface topography, mineral composition, element abundance and soil characteristics to construct a basic geological unit classification dataset with multidimensional features, in order to build a classification model combining machine learning algorithms with feature variables and to conduct research on geological unit classification and prediction. The information used is from the USGS global lunar geological map and the fused data, including the Chang'e-2 CCD camera images and DEM data, the wide angle camera (WAC) data from the Lunar Reconnaissance Orbiter (LROC) and the Kaguya multi-band imager data. This paper first describes a supervised learning method of multi-classification of geological units based on multi-feature variables of data and machine learning algorithms. Then, combined with test results of classification, the algorithm model, the influence of the feature variables and the application of the method are analyzed and discussed. Finally, we summarize the work of this study and directions for future work are provided.

2. Methods

In this study, we extract information of the selected area from multi-source data to form a feature dataset. Through programming via python, we apply a machine learning algorithm to train the data to realize the multi-classification supervised learning and prediction of lunar surface geological units. The overall processes mainly include area selection, feature extraction, geological unit classification and classification result evaluation (Figure 1).

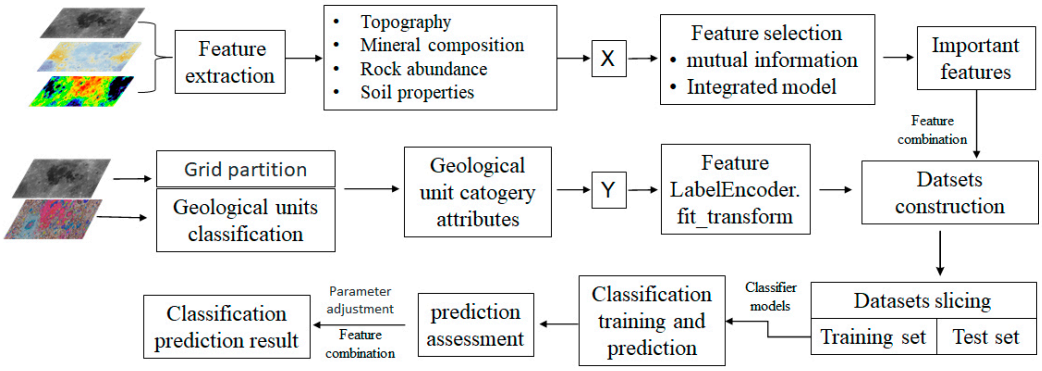

Figure 1. Flow chart of geological unit classification prediction based on machine learning in combination with multi-source data features.

2.1. Study Regions

We selected the Chang'e 5 and Chang'e 4 landing sites as the study area for this work. The Chang'e 5, China's first lunar sample return mission, landed in the northern part of the Oceanus Procellarum and to the west of the Sinus Iridum and Montes Jura [17]. The landing site is flat and ejected materials with high reflectivity can be seen. The authors of [18] indicate that the Chang'e-5 landing site is distributed with multi-period mare basaltic geological units. The Chang'e-4 mission is the first exploration on the lunar far-side. The landing site is located in the Von Kármán crater in the South Pole-Aitken basin [19]. The Von Kármán crater has a flat floor with a prominent central peak, and its overall topography shows a descending trend from northeast to southwest [20]. Some ejected materials from the northeast Finsen crater can be found near the landing site [21]. According to the latest lunar geological map released by USGS [1] (Figure 1c), the Chang'e-5 region contains six types of geological units such as Im2 (Upper Imbrium Mare Unit), Em (Eratosthenian Mare Unit) and Ic1 (Lower Imbirum Crater Unit); and the Chang'e 4 region contains 15 types of geological units such as pNc (pre-Netarian Crater Unit), pNt (pre-Netarian Terra Unit) and Ec (Eratosthenian Crater Unit) (Figure 1c). The topography of the Chang'e-4 landing area is more complex compared to the Chang'e-5 landing area.

In the selected study areas, the Chang'e-2 high-resolution image data are used as the base map for gridding. The image data are divided from a single raster into m x n grid cells according to a certain interval (this can be customized, and an interval of 2 km is used in this study) (Figure 2), and the center pixel of each grid is used as a sample point. The longitude and latitude coordinates of each sample point were calculated to generate the initial vector data of the sample point, which were then used to perform spatial overlay operations with the USGS global geological map vector data to extract the geological unit classification features of each sample point. A total of 23,326 and 18,492 sample points were extracted from the Chang'e 4 and Chang'e 5 landing areas, respectively. The geologic unit classification distribution of each area is shown in Figure 2.

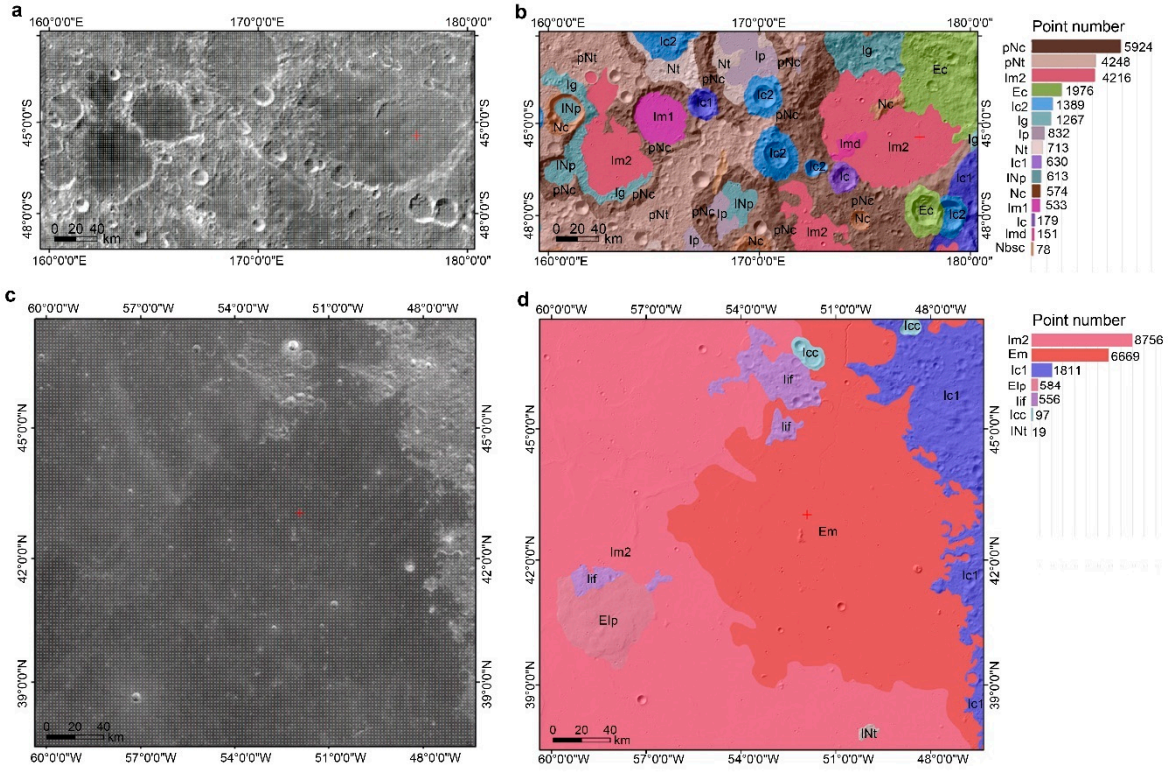

Figure 2. Image and geological unit distribution of the two study areas. (**a**,**b**) Chang'e-4 landing area. (**c**,**d**) Chang'e-5 landing area.

2.2. Feature Extraction

In this study, we extract feature information of data based on Chang'e-2 image and DEM data, wide-angle camera (WAC) data and Kaguya multi-band imager data [22]. Since the grid interval of sampling points is 2 km, the spatial resolution of various data sources should be no less than 2 km to ensure that the feature information corresponding to the sample points can be extracted one-to-one. The spatial resolutions of the selected feature data sources are 20 m, 59 m and 400 m (see Table 1), which are all less than 2 km and meet the requirements. Using the same grid space to perform spatial superposition with various data sources, 14 types of feature information (Table 1) (e.g., topography, mineral abundance, material composition, soil properties) were extracted for each data point according to the pixel position of the sample point, forming the feature samples dataset for geological unit classification.

Table 1. Feature attributes of the sample data and their descriptions.

No	Feature Name	Feature Definition	Data Source Description
1	Longitude	Longitude coordinates of the center of the sample point, ranging from −180 to 180 degrees	The global DOM data were acquired by the Chang'e-2 CCD camera [23,24]. The resolution of the data used is 20 m.
2	Latitude	Latitude coordinates of the center of the sample point, ranging from −90 to 90 degrees	
3	Gray	The grayscale value of the image of the pixel where the sample point is located	
4	Elevation	Elevation from the DEM data of the pixel where the sample point is located	The global DEM data were acquired by the Chang'e-2 CCD camera [24,25]. The resolution of the data used in this paper is 20 m.
5	Relief	The difference between the maximum and minimum elevation values of all pixel points in the eight neighbors centered around the pixel where the sample point is located	
6	slope	The average of the rate of change of elevation from one pixel to another. It can be calculated as $p = atan\left(\sqrt{(\frac{\partial z}{\partial x})^2 + (\frac{\partial z}{\partial y})^2}\right)$, where p is the slope, and $\frac{\partial z}{\partial x}$, $\frac{\partial z}{\partial y}$ denote the partial derivatives in the x and y directions, respectively	
7	TiO2	The TiO2 content of the pixel where the sample point is located	TiO2 content data are from the wide angle camera (WAC) of the US Lunar Reconnaissance Orbiter (LROC) system [26], with data coverage from 0 to 360 degrees longitude and −70 to 70 degrees latitude. The resolution of the data used is 400 m/pixel.
8	FeO	The FeO content of the pixel where the sample point is located	Multispectral image data of the lunar surface acquired by the Kaguya multi-band imager (MI) at five wavelength positions in the ultraviolet-visible band (UVVIS; 415, 750, 900, 950, 1001 nm) and four wavelength positions in the near-infrared band (NIR; 1000, 1050, 1100, 1250 nm). FeO content, four common mineral contents (two types of pyroxene, plagioclase, olivine), submicroscopic metallic iron (SMFe) abundance, and optical maturity (OMAT) data were derived with data coverage from 0 to 360 degrees longitude and −50 to 50 degrees latitude [27]. The resolution of the data used is 59 m/pixel.
9	SMFe	Submicroscopic metallic iron (SMFe) content of the pixel where the sample point is located	
10	Clinopyroxene	Clinopyroxene content of the pixel where the sample point is located	
11	Orthopyroxene	Orthopyroxene content of the pixel where the sample point is located	
12	Plagioclase	Plagioclase content of the pixel where the sample point is located	
13	Olivine	Olivine content of the pixel where the sample point is located	
14	OMAT	Optical maturity of the pixel where the sample point is located	

2.3. Target Classification

The 14 extracted features are used as vector x dataset, and the geological unit classification features are used to construct dataset y as the prediction target of this study. The geological unit classification in dataset y is converted from a non-numerical type to a numerical type by the label encoding (LabelEncoder.fit_transform) method. From the perspective of machine learning, classification can be defined as mapping from one domain (i.e., input data) to another (target classes) via a discrimination function y = f(x). Inputs are represented as m vectors of the form ⟨x1, x2, . . . , xm⟩ and y is a finite set of n class labels {y1, y2, . . . , yn}. Given instances of x and y, supervised machine learning attempts to induce or train a classification model f′, which is an approximation of the discrimination function, $\hat{y} = f'(x)$ and maps input data to target classes [28–30]. In this work, the target classification of geological units consists of three main steps: feature selection, dataset construction and slicing and classification training and prediction.

Feature selection: Feature selection is an important research direction in the field of statistical machine learning, which is central to improving model training speed and classification accuracy and to enhance the interpretability of model results. Too many or too few dimensions of the features, or features without enough importance will eventually, to some extent, lead to the poor generalization of the training model. In this work, a statistical-based estimation of mutual data information and machine-learning-integrated-algorithmic-model-based feature importance evaluation were used for feature selection. The features with higher importance scores were synthetically selected as the preferred features for subsequent feature combination to build the dataset.

Feature dataset construction and slicing: according to the feature importance, the feature variables with significant impact were selected to be combined and to form feature dataset xi. Then, the feature dataset xi and y were randomly sliced into training and test sets simultaneously according to a certain ratio, which is of 70% and 30% in this study. In the specific application, we first trained different models using the training set. Model performance was improved through continuous iterations and the optimal models were selected. Then, we verified and evaluated the performance of the models through the test set.

Classification training and prediction: different classification models were built by combining machine learning classifiers with the dataset for classification training and prediction. In this study, nine machine learning classifiers were first selected. After preliminary testing, KNeighbors, ExtraTree and SVC [31–33], which have poor multi-classification prediction performances, were removed. Six classifiers including DecisionTree, RandomForest, GradientBoosting, XGBoost, CatBoost and Bagging [34–39], which have better performances, were finally selected. The classifiers and the feature dataset were combined to construct classification models for target classification, and their performances were initially judged based on the classification results of the training set. The model was then optimized in two ways; by adjusting the hyperparameters of the algorithm (via grid search or Bayesian algorithm) and by feature selection, to finally make predictions on the test dataset.

2.4. Prediction Assessment

In this study, the target classification of geological units was evaluated using a confusion matrix, accuracy, precision, recall and f1-score. Accuracy is the rate of correct classification of all samples which can better evaluate the overall effectiveness of the model; precision is the correct percentage of all positive predictions; recall is the percentage of correct predictions among all positive samples. This reflects the ability of the algorithm to find all positive samples. A higher value means fewer samples are misclassified; the f1-score is the summed average of precision and recall. An example of the confusion matrix is shown in Figure 3.

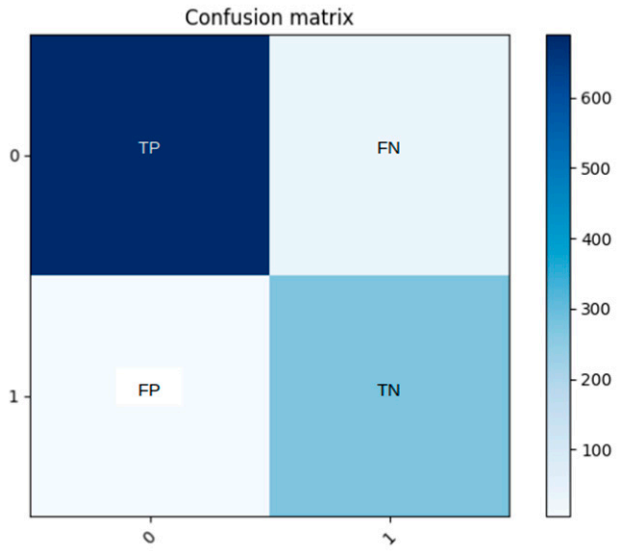

Figure 3. An illustration of Confusion Matrix.

In the confusion matrix, four values are included: (1) True Positive (TP): the number of samples for which both the real label and predicted result are positive; (2) False Negative (FN): the number of samples for which the real label is positive but the predicted result is negative; (3) False Positive (FP): the number of samples for which the real label is negative but the predicted result is positive; (4) True Negative (TN): the number of samples for which both the real label and predicted result are negative. According to these four values, the accuracy, precision, recall and f1-score of the model can be calculated by the following formula:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{f1-score} = \frac{2{*}\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

3. Results

3.1. Constructed Feature Dataset

In this work, two methods were used for feature selection: statistical-based mutual information estimation of data and integrated-algorithmic-model-based evaluation of feature importance. The mutual information estimation based on the nearest neighbor model and the variable importance of the integrated algorithm of XGBoost for feature scoring were applied for the former and latter methods, respectively. The results of feature importance evaluation of both methods are shown in Figure 4. According to the classification result and considering aspects such as topography, geomorphology and mineralogy, eight features with high scores were selected including 'longitude', 'latitude', 'elevation', 'relief', 'TiO2', 'FeO', 'Plagioclase' and 'Olivine'. Feature datasets were then constructed based on these features. Finally, 16 feature datasets (Table 2) were selected for comparison. We combined these feature datasets with the six selected machine learning algorithms to form different classification models for subsequent classification training and testing.

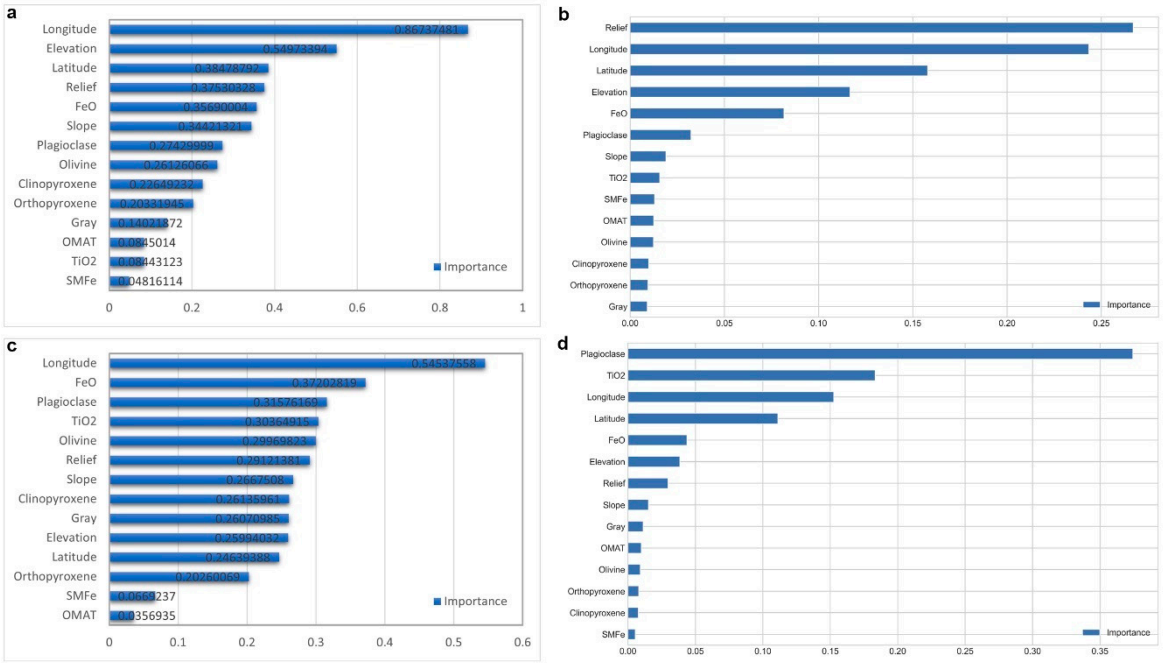

Figure 4. Ranking of feature importance evaluation of two study areas. (**a**,**c**) Scores of feature importance of Chang'e-4 and Chang'e-5 study areas, respectively, based on the nearest neighbor model. (**b**,**d**) Scores of feature importance of Chang'e-4 and Chang'e-5 study areas, respectively, based on the application of the integrated algorithm of XGBoost.

Table 2. Feature dataset formed by combination of important features.

Dataset Name	Abbreviations	Feature Combinations of Datasets
DataSet_1	DS1	'Longitude', 'Latitude', 'FeO'
DataSet_2	DS2	'Longitude', 'Latitude', 'TiO2'
DataSet_3	DS3	'Longitude', 'Latitude', 'Plagioclase'
DataSet_4	DS4	'Longitude', 'Latitude', 'Elevation'
DataSet_5	DS5	'Longitude', 'Latitude', 'Relief'
DataSet_6	DS6	'Longitude', 'Latitude', 'Olivine'
DataSet_7	DS7	'Longitude', 'Latitude', 'Relief', 'TiO2'
DataSet_8	DS8	'Longitude', 'Latitude', 'Elevation', 'TiO2'
DataSet_9	DS9	'Longitude', 'Latitude', 'Relief', 'TiO2', 'Olivine'
DataSet_10	DS10	'Longitude', 'Latitude', 'Elevation', 'TiO2', 'Olivine'
DataSet_11	DS11	'Longitude', 'Latitude', 'Relief', 'TiO2', 'FeO'
DataSet_12	DS12	'Longitude', 'Latitude', 'Elevation', 'TiO2', 'FeO'
DataSet_13	DS13	'Longitude', 'Latitude', 'Relief', 'TiO2', 'Plagioclase'
DataSet_14	DS14	'Longitude', 'Latitude', 'Elevation', 'TiO2', 'Plagioclase'
DataSet_15	DS15	'Longitude', 'Latitude', 'Relief', 'TiO2', 'Plagioclase', 'FeO'
DataSet_16	DS16	'Longitude', 'Latitude', 'Elevation', 'TiO2', 'FeO', 'Plagioclase'

3.2. Chang'e-4 Landing Area

The geological background of the Chang'e 4 landing area is very complex, and there are as many as 15 types of geological unit classifications. The average classification accuracy of the six algorithms on the 15 datasets is higher than 89.5%. The algorithms ranked from highest to lowest in terms of classification ability are: XGBoost, Bagging, CatBoost, GradientBoosting, DecisionTree and RandomForest (Figure 5a). The highest classification accuracy of 95.1% is obtained by XGBoost + DataSet_4; the following highest classification accuracy is obtained by XGBoost on the datasets of DataSet_8, DataSet_14 and DataSet_10 with accuracies of 95%, by Bagging + DataSet_2 with accuracies of 94.9% and by CatBoost + DataSet_4 with accuracies of 94.8%. The highest value of f1-score is 93.8%, which is

obtained by Bagging + DataSet_2; the second value is 92.9%, which is obtained by XGBoost + DataSet_4 and Bagging + DataSet_4 (Figure 5d). It can be seen that the algorithm with the strongest classification ability in this region is still XGBoost, and the most effective feature sets are DataSet_4, DataSet_8, DataSet_2, DataSet_14 and DataSet_10. According to the statistics with maximum range, except for latitude and longitude, other features involved in these datasets are 'elevation', 'TiO2', 'Plagioclase' and 'Olivine'.

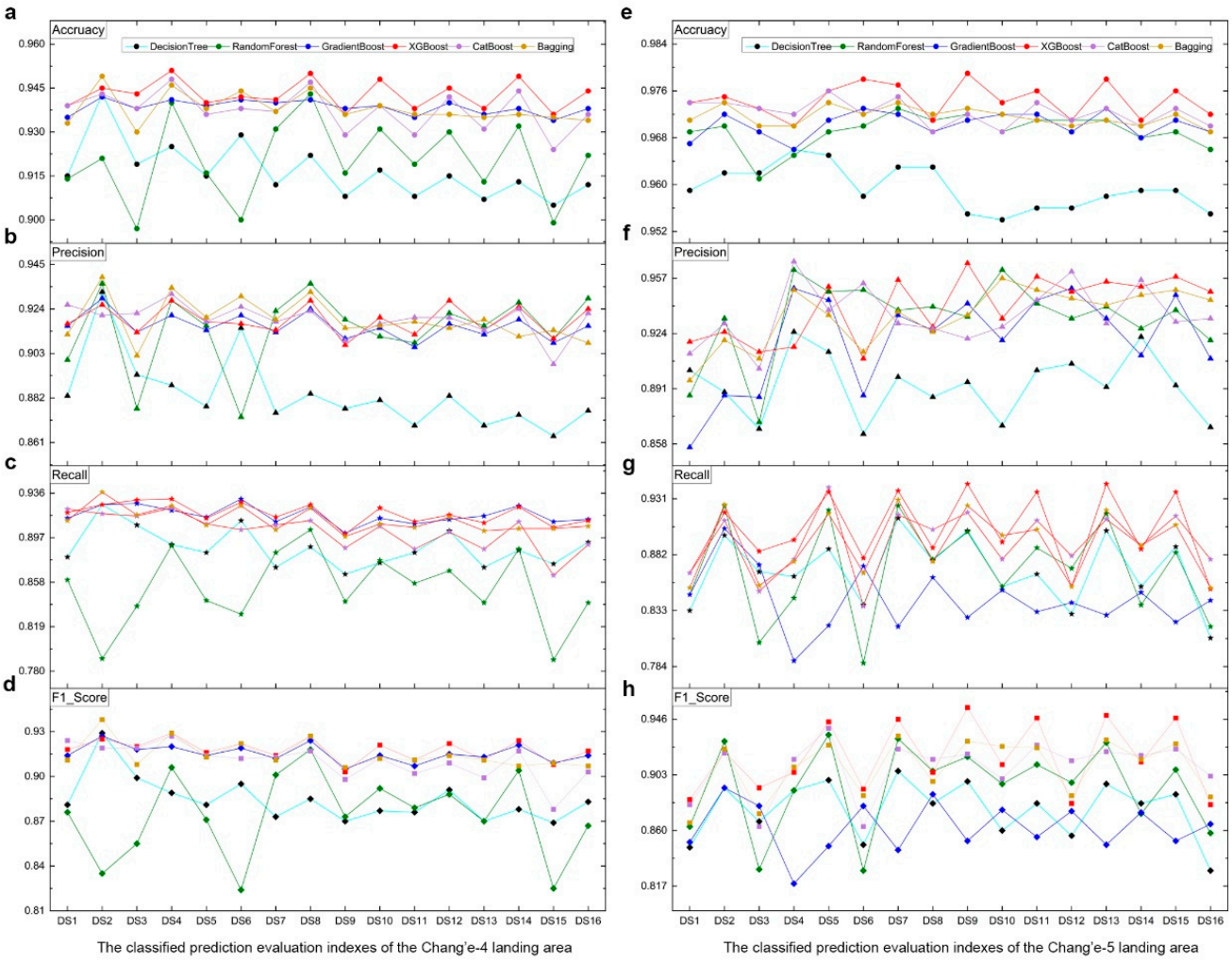

Figure 5. The summary of obtained accuracy, precision, recall and f1-score for each classification model of the two study areas. (**a**–**d**) The landing area of Chang'e-4, and (**e**–**h**) the landing area of Chang'e-5.

According to further analysis of the classification evaluation report of the classification model (XGBoost + DataSet_4) with the highest accuracy score in this study area (Figure 6a), the algorithmic model exhibits a strong classification ability in the classification prediction for almost all the geological units, and only has a slightly lower classification prediction ability in the Nbsc geological unit which is represented by 10 with a lower number of samples (Figure 6b). A total of 345 samples are misclassified out of 6997 of the overall test samples, and the overall classification accuracy of the model reaches a high level with an micro-average f1-score of 0.929 for all types of predictions.

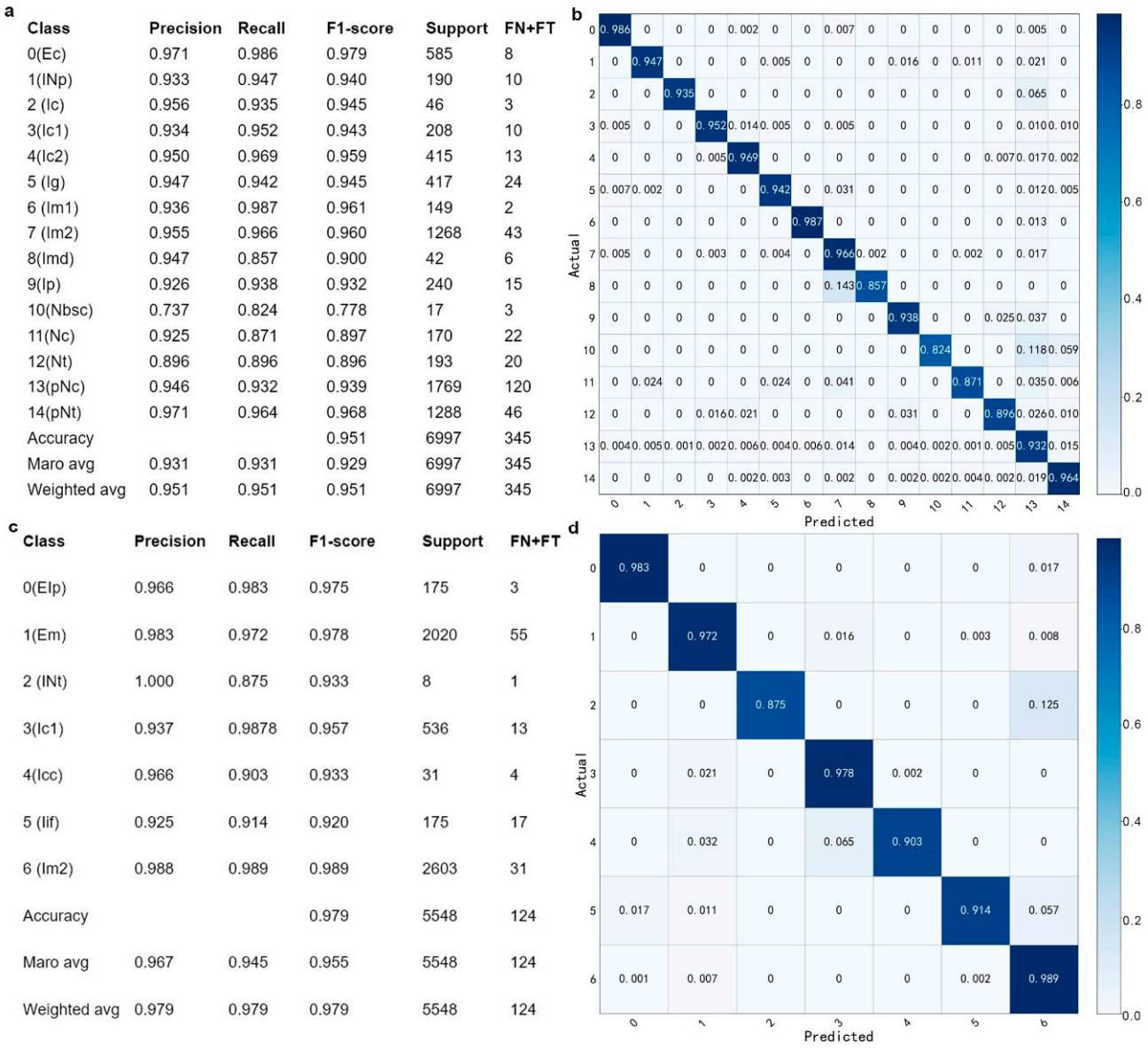

Figure 6. Classification evaluation report and the Recall–Confusion Matrix of the highest classification accuracy for the two study areas. (**a**,**b**) Classification results of Chang'e-4 landing area obtained on XGBoost + DataSet_4. (**c**,**d**) Classification results of Chang'e-5 landing area obtained on XGBoost + DataSet_9.

3.3. Chang'e-5 Landing Area

Compared to Chang'e-4, the geological background of the Chang'e-5 landing area is relatively simple. Within the Chang'e-5 study area, all the six algorithms on the 15 datasets reached a classification accuracy of 95.3% or higher. The algorithms ranked from highest to lowest in terms of classification prediction ability are: XGBoost, CatBoost, Bagging, GradientBoosting, RandomForest and DecisionTree. The highest classification accuracy is 97.9%, which is obtained by the combination of XGBoost + DataSet_9; the second highest values are 97.8% and 97.7%, which are obtained by XGBoost on the datasets of DataSet_6, DataSet_13 and DataSet_7 (Figure 5e). The highest value of f1-score is 95.5%, which is obtained by XGBoost + DataSet_9. The second highest values are 94.9% and 94.7%, which are obtained by XGBoost on the datasets of DataSet_13, DataSet_11 and DataSet_15 (Figure 5h). It can be seen that the algorithm with the strongest classification ability in this region is XGBoost, and the most effective feature sets are DataSet_9, DataSet_13, DataSet_6,

DataSet_7, DataSet_11 and DataSet_15. According to the statistics with maximum range, except for latitude and longitude, the features involved by these datasets are 'relief', 'TiO2', 'Plagioclase', 'Olivine' and 'FeO'.

According to further analysis of the classification evaluation report of XGBoost + DataSet_9, which obtained the highest classification accuracy score in this study area (Figure 6c), the model shows excellent classification performance on the prediction of the Im2 and Em geologic units represented by 6 and 1, respectively, which have the largest number of samples. The model also achieves high prediction accuracy and recall rates on the INt and Icc geologic units represented by 2 and 4, respectively, which have a lower number of samples (Figure 6d). A total of 122 samples are misclassified out of the overall 5548 test samples, and the overall classification accuracy of the model is high with a macro-average f1-score of 0.949 for all types of predictions.

4. Discussions

4.1. Comparison of Classification Models

In this study, we proposed a classification model constructed for the classification prediction of geological units by combing machine learning classification algorithms with features of data. High accuracy classification results of 97.8% and 95.1% were obtained for the two study areas, which demonstrates the effectiveness of the classification models. Firstly, according to the test results of the two study areas, all the six machine learning classification algorithms exhibit a strong classification ability. The algorithms of XGBoost, CatBoost, Bagging and GradientBoosting are better than RandomForest and DecisionTree in terms of classification ability (Figure 7a,c). XGBoost is the best algorithm which obtains the highest classification prediction accuracy and comprehensive average classification accuracy for both study regions. Moreover, the lowest value of classification accuracy is also higher than other algorithms, indicating that XGBoost has high stability. In this study, we improved the model in two aspects according to the training results. First, we used the Bayesian optimization algorithm to optimize and adjust the parameters of the XGBoost algorithm. Taking into consideration operational efficiency, the parameters were finally determined as follows: learning_rate = 0.01, n_estimators = 1000, max_depth = 10, min_child_weight = 1, gamma = 0, subsample = 1, colsample_bytree = 1, objective= 'multi:softmax' and seed = 1. The number of iterations (n_estimators) for the CatBoost, Bagging and GradientBoosting classification algorithms were also taken as 1000 times. Secondly, the prediction results also show that the same machine learning algorithm has a significant difference in classification ability on different feature datasets (Figure 7b,d). These datasets were constructed by combining features indicating the significant impact of the combination of data features on the classification results. Therefore, we focus on testing the features that exhibit a significant influence in training, and select the most effective feature combination to form a feature set through a large number of experiments and combine it with the machine learning algorithm to build a classification model with higher accuracy. Through this test, we found that it is a complex task of lesser effectiveness to improve the performance of the machine learning algorithm by adjusting the hyperparameters of the algorithm in order to achieve a higher classification prediction accuracy. However, by combining features to build different datasets, we can achieve significantly different prediction accuracies. Compared to adjusting the algorithm parameters, building an effective feature dataset by selecting and combining features is a more effective way to improve the classification prediction accuracy.

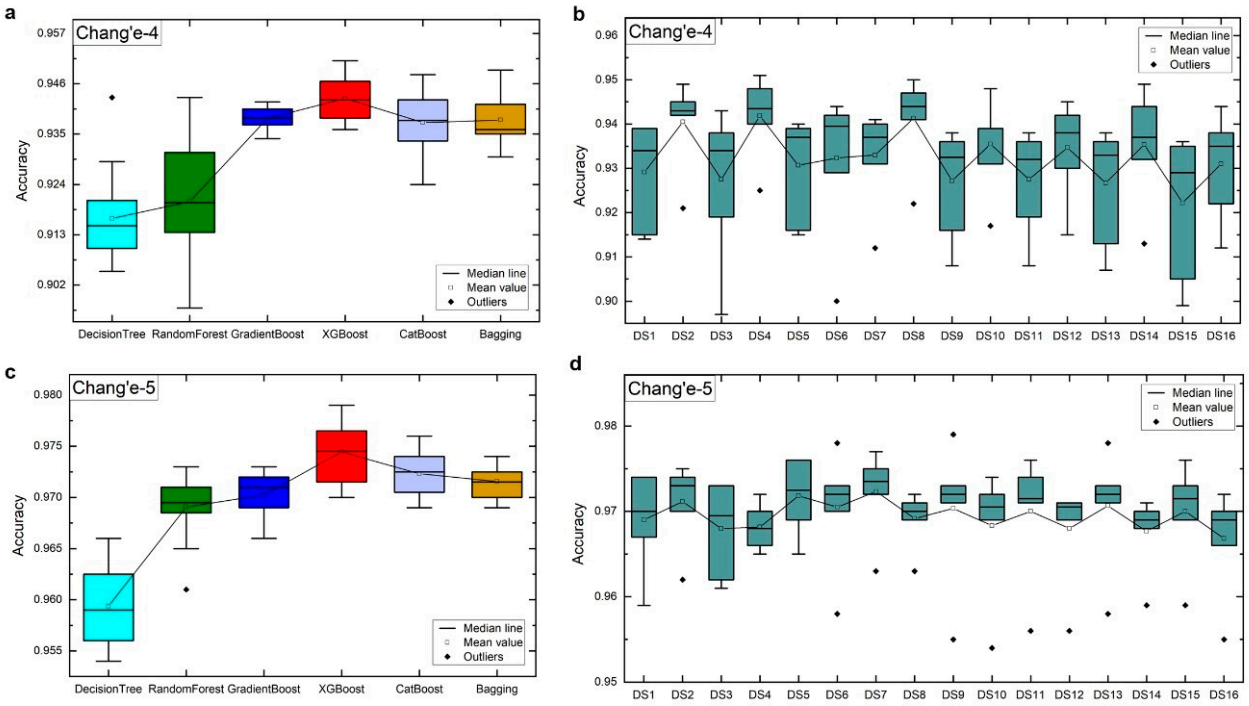

Figure 7. Statistics of the means of the accuracy for 6 machine learning algorithms and 16 datasets. (**a**,**b**) Chang'e-4 area. (**c**,**d**) Chang'e-5 area.

4.2. Feature Selection and Correlation Analysis

In this study, we selected eight important features including 'longitude', 'latitude', 'elevation', 'relief', 'TiO2', 'FeO', 'Plagioclase' and 'Olivine'. Through the combination of these features, the high accuracy of the classification prediction of geological units were acquired for the two study areas, verifying the effectiveness of the feature selection method. To distinguish the differences of the geological units in each region (e.g., Figure 8) in terms of morphological and tectonic characteristics, the feature with the most significant influence is 'elevation' for the Chang'e-4 region and 'relief' for the Chang'e-5 region. In terms of material composition and mineral abundance, the features with the most significant influence in both regions are 'Olivine', 'TiO2' and 'Plagioclase'. The feature 'FeO' obtains a high score in the importance assessment of feature selection, but the measured classification accuracy influence is slightly less than the former features. Comparing the top three classification prediction results of the two regions, the influence of these features is higher on the Chang'e-5 region than on the Chang'e-4 region. The 'Olivine' features are negatively correlated with 'elevation' and 'relief' for Chang'e-4 and Chang'e-5, respectively. Compared to the surrounding geological units, a higher olivine content can be seen in the units (Im1, Im2, Imd) with lower elevation in the Chang'e-4 region and in the units (Em, Im2) with lower undulation in the Chang'e 5 region. Combined with the feature analysis of the Chang'e-5 area, the older geological units (Ic1, Icc, Iif, INt) have a higher plagioclase content than the younger ones (Em), while the younger geological units (Em) have a higher TiO2 content than the older ones (Ic1, Icc, Iif, INt). These findings are consistent with the results of existing studies [18]. However, these two findings are not observed for the Chang'e-4 region. This could be due to the fact that the materials of the geological units in the Chang'e-4 region are mixed with the ejected materials from the surrounding highlands, which results in the material composition of the region being more difficult to distinguish.

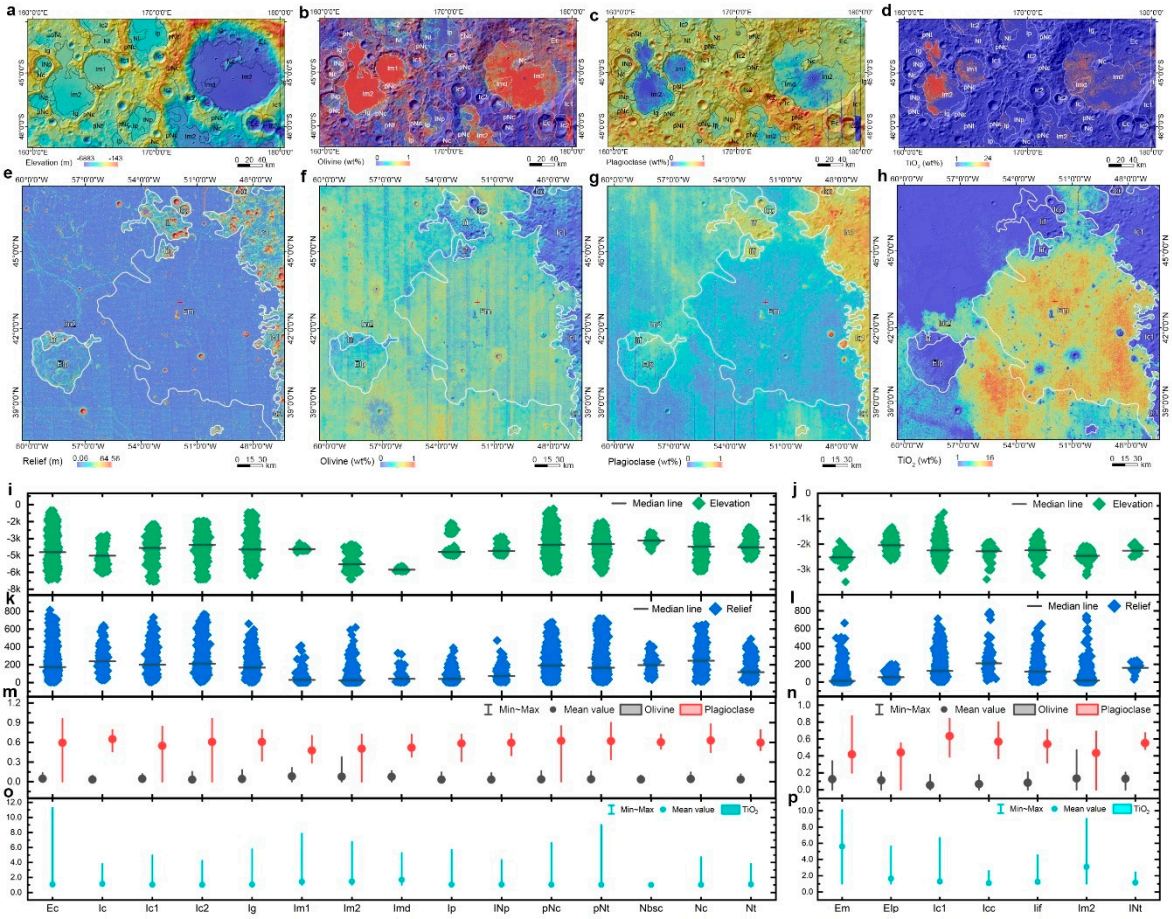

Figure 8. Feature distribution and statistics of featured geological unit classification. (**a**) Topographic elevation distribution of CE-4 landing area. (**b**) Olivine abundance of CE-4 landing area. (**c**) Plagioclase content of CE-4 landing area. (**d**) TiO2 content of CE-4 landing area. (**e**) Relief of CE-5 landing area. (**f**) Olivine content of CE-5 landing area. (**g**) Plagioclase content of CE-5 landing area. (**h**) TiO2 distribution of CE-5 landing area. (**i**) Statistic of elevation of classified geological units for CE-4 landing area. (**j**) Statistic of elevation of classified geological units for CE-5 landing area. (**k**) Statistic of relief of classified geological units for CE-4 landing area. (**l**) Statistic of relief of classified geological units for CE-5 landing area. (**m**) Statistic of olivine and plagioclase of classified geological units for CE-4 landing area. (**n**) Statistic of olivine and plagioclase of classified geological units for CE-5 landing area. (**o**) Statistic of TiO2 of classified geological units for CE-4 landing area. (**p**) Statistic of TiO2 of classified geological units for CE-5 landing area.

4.3. Applications of Classification Prediction

The classification prediction method of lunar surface geological units proposed in this study can be used for the geological unit classifying and mapping of global lunar digital mapping. Based on 70% of the known information about the study areas, the method effectively achieves 95.1% and 97.9% of geological unit classification for the study areas of Chang'e-4 and Chang'e-5, respectively. Based on 50% of the known information about the study areas, the method can achieve 93.9% and 97.3% of geological unit classification for the study areas of Chang'e-4 and Chang'e-5, respectively (Table 3). From the prediction results (Figure 9), the method still has a high identification capacity of complex geological units (e.g., as many as 15 geological units in the Chang'e-4 area) and for geological units accounting for a relatively low percentage of the investigated area (e.g., Imd, Ic and Nbsc geological units represented by 2, 6 and 8 in CE-4 landing areas, respectively, as shown in Figure 6a; INt and Icc geological units represented by 2 and 4 in CE-5 landing areas, respectively, as shown in Figure 6c). Even in scenarios wherein the distribution of geological units is very complex

and multiple types of geological units occur alternatively, the corresponding geological unit boundaries can still be effectively delineated. The overall identification accuracy of the method is high with a good mapping result. It can also be seen from Figure 9 that most of the sample points with more identification errors occur at the boundaries between geological units. The identification accuracy can be enhanced in terms of two aspects, including improving the delineation precision of the pixels and selecting and combining features in a specific manner by conducting geological surveys in specific regions. These two aspects can ensure a higher accuracy in the classification and prediction.

Table 3. Results of classification tests with different allocation ratios of known and unknown information.

Study Regions	Classification Model	Number of Training Samples	Number of Test Samples	Proportion of Known Information	Accuracy
CE-4	XGBoost + DataSet_4	16326	6997	70%	95.1%
		13993	9330	60%	94.6%
		11662	11662	50%	93.9%
CE-5	XGBoost + DataSet_9	12944	5548	70%	97.9%
		11095	7397	60%	97.6%
		9246	9246	50%	97.3%

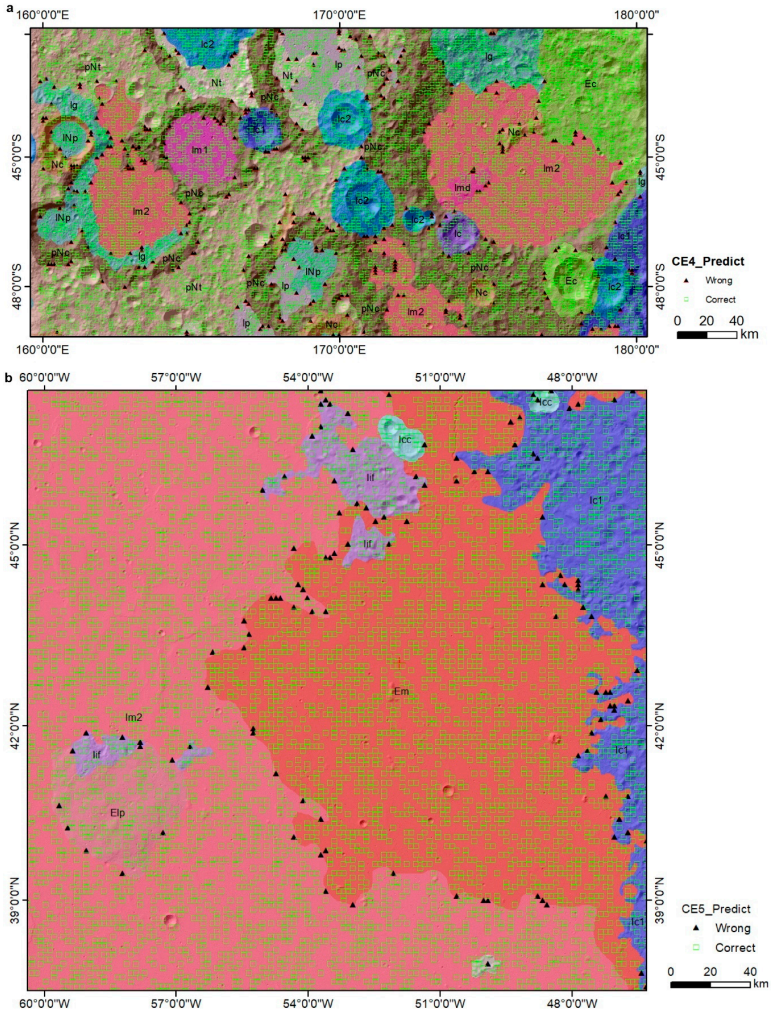

Figure 9. Visualization of classification prediction results of geological units for two study areas. (**a**) Chang'e-4 landing area. (**b**) Chang'e-5 landing area.

5. Conclusions

In this study, we develop a method of combining machine learning algorithms with features of data to build a classification model for the classification prediction of geological units. Using the Chang'e-4 and Chang'e-5 landing sites as our testing areas, we verify that the developed method is a useful exploration and practice on the classification prediction of geological units on the lunar surface. The main findings are as follows:

(1) Classification models: the classification models constructed obtain high accuracy classification predictions of 97.9% and 95.1% for the two inhomogeneous and complex areas with multiple classifications. This fully verifies the effectiveness of the constructed classification models, which combine machine learning algorithms with data features (e.g., topography, geomorphology, mineral abundance, material composition) in the classification prediction of geological units. On one hand, all the six machine learning algorithms selected exhibit a strong multi-classification ability, among which XGBoost, CatBoost, Bagging and GradientBoosting are preferred, and especially XGBoost which has the best classification performance and can be used as the preferred classifier for subsequent work; on the other hand, the feature dataset composed of the combination of feature variables has an important influence on the accuracy of geological unit classification prediction. Compared to adjusting the hyperparameters of the machine learning algorithm, building an effective feature dataset by feature combination is a more effective way to improve the classification prediction accuracy.

(2) Feature selection: several important features such as 'elevation', 'relief', 'TiO2', 'Plagioclase', 'Olivine' and 'FeO' were screened using the two feature selection methods, namely, statistical-based data mutual information estimation and model-based machine learning algorithm feature evaluation. A classification model was constructed by the combination of these features to achieve a high accuracy geological unit classification prediction. These features also effectively reflect the apparent variation in topography, geomorphology, materials composition and mineral abundance of the study areas, which deepens our understandings on the formation and evolution of the Moon. It should be noted that although the final classification prediction results verify the effectiveness of the feature selection method, the features selected in this study are not the only features that can be used due to the diversity of the feature selection methods. The effectiveness of other features and their associated combinations is still worth exploring. Therefore, our future work will focus on mining more effective feature variables to obtain more accurate classification prediction results and conducting in-depth research on correlation analysis between data features and geological units.

(3) Application of the method: The developed method is flexible, efficient and has a good extensibility. It is suitable for the geological unit classification prediction for any lunar geological map data and any region of the Moon. The classification prediction method can not only be applied to the digital mapping of the global Moon surface, but also provide effective support for the automatic mapping of geological units in any region. In addition, effective feature variables can be mined through classification prediction, which can help to perform in-depth comprehensive analysis of geological units for any size area on the lunar surface. Moreover, the classification method can also be applied to the classification of lunar surface chronological units. Subsequently, we will attempt to mine the association rules of geochronological units on the global lunar surface based on the results of this work. A lunar surface chronology and quantitative analysis model based on machine learning of multiple feature variables will be also our central focus in the future.

Author Contributions: Conceptualization, W.Z. and C.L.; Data curation, X.Z., X.G. and Z.Z.; Formal analysis, W.Z. and X.Z.; Methodology, W.Z.; Resources, X.Z., X.G. and D.L.; Software, W.Z. and X.G.; Validation, X.Z. and X.G.; Visualization, W.Z. and X.Z.; Writing—Original draft, W.Z., X.Z. and X.G.; Writing—Review and editing, Z.Z., D.L. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research program of Chinese Academy of Sciences (ZDBS-SSW-JSC007).

Data Availability Statement: The multi-source data of feature extracted in this study is available at the following address: Chang'e-2 DOM data: http://www.dx.doi.org/10.12350/CLPDS.GRAS.CE2.DOM-20m.vA (accessed on 6 February 2012); Chang'e-2 DEM data: http://www.dx.doi.org/10.12350/CLPDS.GRAS.CE2.DEM-20m.vA (accessed on 6 February 2012); TiO2 content data: https://wms.lroc.asu.edu/lroc/view_rdr/WAC_TIO2 (accessed on 1 November 2017); FeO contents data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_FeOWeightPercent_50N50S (accessed on 1 September 2016); Submicroscopic metallic iron (SMFe) abundance data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_AbundanceSMFe_50N50S?p=2&pb=1#downloads (accessed on 1 September 2016); Clinopyroxene contents data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_ClinopyroxenePercent_50N50S (accessed on 1 September 2016); Orthopyroxene contents data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_OrthopyroxenePercent_50N50S (accessed on 1 September 2016); Plagioclase contents data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_PlagioclasePercent_50N50S (accessed on 1 September 2016); Olivine content data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_OlivinePercent_50N50S (accessed on 1 September 2016); Optical maturity (OMAT) data: https://astrogeology.usgs.gov/search/map/Moon/Kaguya/MI/MineralMaps/Lunar_Kaguya_MIMap_MineralDeconv_OpticalMaturityIndex_50N50S (accessed on 1 September 2016).

Acknowledgments: We thank the team members of the Ground Research and Application System (GRAS) who have contributed to the Chang'e project data receiving, preprocessing, management and release.

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

References

1. Fortezzo, C.M.; Spudis, P.D.; Harrel, S.L. Release of the digital unified global geologic map of the Moon at 1: 5,000,000-Scale. *Lunar Planet. Sci. Conf.* **2020**, *2326*, 2760.
2. Ouyang, Z.Y.; Liu, J.Z. The origin and evolution of the Moon and its geological mapping. *Earth Sci. Front.* **2014**, *21*, 1–6. (In Chinese)
3. Ling, Z.C.; Liu, J.Z.; Zhang, J.; Li, B.; Wu, Z.C.; Ni, Y.H.; Sun, L.Z. The lunar rock types as determined by Chang'E-1 IIM data: A case study of Mare Imbrium-Mare Frigoris region (LQ-4). *Adv. Earth Sci.* **2014**, *21*, 107–120. (In Chinese)
4. Ding, X.Z.; Wang, L.; Han, K.Y.; Pang, J.F.; Liu, J.Z.; Guo, D.J.; Ding, W.W.; Ju, Y.J. The lunar digital geological mapping based on ArcGIS: Taking the arctic region as an example. *Adv. Earth Sci.* **2014**, *21*, 19–30. (In Chinese)
5. Cheng, W.M.; Liu, Q.Y.; Wang, J.; Gao, W.X.; Liu, J.Z. A preliminary study of classification method on lunar topography and landforms. *Adv. Earth Sci.* **2018**, *33*, 885–897. (In Chinese)
6. Cracknell, M.J.; Reading, A.M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput. Geosci.* **2014**, *63*, 22–33. [CrossRef]
7. Harris, J.R.; Grunsky, E.C. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Comput. Geosci.* **2015**, *80*, 9–25. [CrossRef]
8. Zheng, Y. Research on Lithology Recognition Based on Deep Learning. Ph.D. Thesis, China University of Petroleum, Beijing, China, 2017. (In Chinese with English abstract).
9. Othman, A.A.; Gloaguen, R. Integration of spectral, spatial and morphometric data into lithological mapping: A comparison of different Machine Learning Algorithms in the Kurdistan Region, NE Iraq. *J. Asian Earth Sci.* **2017**, *146*, 90–102. [CrossRef]
10. Kuhn, S.; Cracknell, M.J.; Reading, A.M. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. *Geophysics* **2018**, *83*, B183–B193. [CrossRef]
11. Zhang, Y.; Sun, J.; Yu, C.C.; Meng, P.Y.; Guo, Z. Classification of quaternary coverings in desert grassland shallow cover area based on multi-source remote sensing data: A case of 1:50000 pilot geological mapping in Qigandiani, Inner Mongolia. *Bull. Geol. Sci. Technol.* **2019**, *38*, 281–290. (In Chinese)
12. Duan, Y.X.; Zhao, Y.S.; Ma, C.F.; Jiang, W.X. Lithology identification method based on multi-layer ensemble learning. *J. Data Acquis. Process.* **2020**, *35*, 572–581. (In Chinese)
13. Zhu, M.Y.; Li, B.Q.; Fu, H.Z.; Chen, C.; Gao, M. SVM lithological classification based on multi-source data collaboration: A case study in Jianggalesayi area. *Uranium Geol.* **2020**, *36*, 288–292+317. (In Chinese)

14. Wang, J. Mineral Assemblages Mapping of Porphyry Copper Deposits Based on Normalized Multispectral Remote Sensing Data in the Dulong Ore Concentrating Area. Ph.D. Thesis, Chengdu University of Technology, Chengdu, China, 2018. (In Chinese with English abstract)
15. Wang, Q.; Lin, B.; Tang, J.X.; Song, Y.; Li, Y.B.; Hou, J.F.; Wei, L.J. Diagenesis, lithogenesis and geodynamic setting of intrusions in Senadong Area, Duolong district, Tibet. *Earth Sci.* **2018**, *43*, 1125–1141.
16. Wu, G.; Chen, G.; Cheng, Q.; Zhang, Z.; Yang, J. Unsupervised machine learning for lithological mapping using geochemical data in covered areas of Jining, China. *Nat. Resour. Res.* **2021**, *30*, 1053–1068. [CrossRef]
17. Li, C.; Hu, H.; Yang, M.F.; Pei, Z.Y.; Zhou, Q.; Ren, X.; Liu, B.; Liu, D.; Zeng, X.; Zhang, G.; et al. Characteristics of the lunar samples returned by the Chang'E-5 mission. *Natl. Sci. Rev.* **2022**, *9*, nwab188. [CrossRef]
18. Qian, Y.Q.; Xiao, L.; Zhao, S.Y. Geology and scientific significance of the Rümker region in northern Oceanus Procellarum: China's Chang'E-5 landing region. *J. Geophys. Res. Planets* **2018**, *123*, 1407–1430. [CrossRef]
19. Liu, J.; Ren, X.; Yan, W. Descent trajectory reconstruction and landing site positioning of Chang'E-4 on the lunar farside. *Nat. Commun.* **2019**, *10*, 4229. [CrossRef]
20. Di, K.; Liu, Z.; Liu, B.; Wan, W.; Peng, M.; Wang, Y.; Gou, S.; Yue, Z.; Xin, X.; Jia, M.; et al. Chang'e-4 lander localization based on multi-source data. *J. Remote Sens* **2019**, *23*, 177–184.
21. Li, C.; Liu, D.; Liu, B.; Ren, X.; Liu, J.; He, Z.; Zuo, W.; Zeng, X.; Xu, R.; Tan, X.; et al. Chang'E-4 initial spectroscopic identification of lunar far-side mantle-derived materials. *Nature* **2019**, *569*, 378–382. [CrossRef]
22. Ohtake, M.; Pieters, C.M.; Isaacson, P.; Besse, S.; Yokota, Y.; Matsunaga, T.; Boardman, J.; Yamomoto, S.; Haruyama, J.; Staid, M.; et al. One Moon, many measurements 3: Spectral reflectance. *Icarus* **2013**, *226*, 364–374. [CrossRef]
23. Li, C.L.; Liu, J.J.; Ren, X.; Mou, L.L.; Mou, L.L.; Zou, Y.L.; Zhang, H.B.; Lü, C.; Liu, J.Z.; Zuo, W.; et al. The global image of the moon by the Chang'E-1: Data processing and lunar cartography. *Sci. China Earth Sci.* **2010**, *53*, 1091–1102. [CrossRef]
24. Zuo, W.; Li, C.; Zhang, Z.; Zeng, X.; Liu, Y.; Xiong, Y. China's Lunar and Planetary Data System: Preserve and Present Reliable Chang'e Project and Tianwen-1 Scientific. *Space Sci. Rev.* **2021**, *217*, 88. [CrossRef]
25. Li, C.; Ren, X.; Liu, J.; Zou, X.; Mu, L.; Wang, J.; Shu, R.; Zou, Y.; Zhang, H.; Lü, C.; et al. Laser altimetry data of Chang'E-1 and the global lunar DEM model. *Sci. China Earth Sci.* **2010**, *53*, 1582–1593. [CrossRef]
26. Sato, H.; Robinson, M.S.; Lawrence, S.J.; Denevi, B.W.; Hapke, H.; Jolliff, B.L.; Hiesinger, H. Lunar Mare TiO2 Abundances Estimated from UV/Vis Reflectance. *Icarus* **2017**, *296*, 216–238. [CrossRef]
27. Lemelin, M.; Lucey, P.G.; Song, E.; Taylor, G.J. Lunar central peak mineralogy and iron content using the Kaguya Multiband Imager: Reassessment of the compositional structure of the lunar crust. *J. Geophys. Res. Planets* **2015**, *120*, 869–887. [CrossRef]
28. Gahegan, M. On the application of inductive machine learning tools to geographical analysis. *Geogr. Anal.* **2000**, *32*, 113–139. [CrossRef]
29. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
30. Kovacevic, M.; Bajat, B.; Trivic, B.; Pavlovic, R. Geological units classification of multispectral images by using Support Vector Machines. In Proceedings of the International Conference on Intelligent Networking and Collaborative Systems, Barcelona, Spain, 4–6 November 2009; pp. 267–272.
31. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.
32. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
33. Hsu, C.W.; Chang, C.C.; Lin, C.J. *A Practical Guide to Support Vector Classification*; Department of Computer Science, Mational Taiwan University: Taipei, Taiwan, 2003.
34. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
35. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
36. Friedman, J.H. Greedy function approximation, a gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
37. Chen, T.; Guestrin, C. Xgboost, A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
38. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363, 2018.
39. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]