International Journal of Geo-Information

MDPI

Article

Efficiency of Extreme Gradient Boosting for Imbalanced Land Cover Classification Using an Extended Margin and Disagreement Performance

Fei Sun [1], Run Wang [1,2], Bo Wan [1,2,*], Yanjun Su [3], Qinghua Guo [3], Youxin Huang [1] and Xincai Wu [1]

1 School of Geography and of Information Engineering, China University of Geosciences, No. 388 Lumo Road, Wuhan 430074, China
2 National Engineering Research Center of Geographic Information System, Wuhan 430074, China
3 State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
* Correspondence: wanbo@cug.edu.cn; Tel.: +86-027-6788-3728

Received: 31 May 2019; Accepted: 20 July 2019; Published: 23 July 2019

Abstract: Imbalanced learning is a methodological challenge in remote sensing communities, especially in complex areas where the spectral similarity exists between land covers. Obtaining high-confidence classification results for imbalanced class issues is highly important in practice. In this paper, extreme gradient boosting (XGB), a novel tree-based ensemble system, is employed to classify the land cover types in Very-high resolution (VHR) images with imbalanced training data. We introduce an extended margin criterion and disagreement performance to evaluate the efficiency of XGB in imbalanced learning situations and examine the effect of minority class spectral separability on model performance. The results suggest that the uncertainty of XGB associated with correct classification is stable. The average probability-based margin of correct classification provided by XGB is 0.82, which is about 46.30% higher than that by random forest (RF) method (0.56). Moreover, the performance uncertainty of XGB is insensitive to spectral separability after the sample imbalance reached a certain level (minority:majority > 10:100). The impact of sample imbalance on the minority class is also related to its spectral separability, and XGB performs better than RF in terms of user accuracy for the minority class with imperfect separability. The disagreement components of XGB are better and more stable than RF with imbalanced samples, especially for complex areas with more types. In addition, appropriate sample imbalance helps to improve the trade-off between the recognition accuracy of XGB and the sample cost. According to our analysis, this margin-based uncertainty assessment and disagreement performance can help users identify the confidence level and error component in similar classification performance (overall, producer, and user accuracies).

Keywords: gradient boosting trees; margin; class imbalance; very-high resolution (VHR) remote sensing; land cover classification; disagreement performance

1. Introduction

Accurate spatially explicit thematic maps with high confidence levels derived from remote sensing images are an important information resource in earth science applications [1–3]. Due to the lack of ancillary information and the high costs of obtaining labelled data [4], land agencies must make decisions using thematic classification maps trained, calibrated and validated using limited reference data [1,5]. These limited data cause an imbalance in the class distribution of training datasets, potentially leading to reduced classification accuracy and uncertainty [1,6]. Since training data are the basis upon which supervised classifiers are constructed [7], the imbalance problem has a huge impact

on most machine learning and classification methods, which assume that the training data are balanced and have the same distribution as the full dataset [8,9]. Generally, however, this assumption cannot be satisfied in regional VHR image land cover classification (LCC) scenarios. Furthermore, the validation dataset is also assumed to have the same distribution as the training dataset [1]; however, the real conditions are usually unknown.

Technically, any dataset suffering from an unequal class distribution will result in poor accuracy for minority classes but in favour of the majority class [10]. Imbalance issues are common in the machine learning (ML) and remote sensing fields [1,11]. Technologies that capture recent advances in the imbalanced learning issues have been published in the field of machine learning (ML). They can be divided into 3 categories, internal approaches acting on the algorithm, on the data or the combined approaches that are based on ensembles of classifiers [12]. At algorithm level, the most common procedure is combing with cost-sensitive learning [13]. For preprocessing imbalanced data set, down-sampling [11,14], over-sampling [15], combination of the two or synthetic technique [16], and join different weight when sampling have also been explored [17]. Adopting diverse classification strategies is also effective [12,18]. The applications employing these technologies to solve the realistic imbalance problems involve various fields. For example, Krawczyk [19] and Hassan [20] adopted ensemble learning and under-sampling for breast cancer malignancy grading and automobile insurance fraud detection, respectively. López [21] used cost-sensitive and fuzzy logic combining for big data classification. Wu [22] optimized the ELM algorithm with mixed-kernel weights for classification with imbalance human activity data.

However, from the remote sensing perspective, imbalanced learning problems are discussed from different aspects. A two-phase technique for the multilayer perceptron (MLP) was presented to speed up training with imbalanced data [23], and a cost-effective network extension scheme for the convolutional neural network (CNN) was introduced to categorize imbalanced aerial images [24]. A semi-supervised technology with unlabelled samples was used to engineer features for recognizing data with skewed distributions [25]. Studies have also explored the influences of imbalanced samples. A new ensemble margin criterion was introduced to evaluate the uncertainty of an RF when performing imbalanced-forest classification [1], and the effect of imbalanced remote sensing data on a support vector machine model was also quantified [26]. However, practitioners will encounter different problems when using the existing techniques. An appropriate cost matrix is difficult to find when performing cost-sensitive learning [13]. Moreover, selection of resampling strategies also depends on many factors [27]. For neural networks (including CNN), which are black-box models, large amounts of computation are usually required [24].

Models that are free from statistical assumptions are desirable in imbalanced LCC scenarios from VHR images. Therefore, decision tree (DT), a classical data-derived rule [12] model is considered [28,29]. Extreme gradient boosting (XGB) [30] is a state-of-the-art gradient boosting ensemble tree model that has recently been widely used [31,32]. It is a scalable, regularized gradient boosting technology that provides predictive performance, especially for high-dimensional problems. As an ensemble tree model, XGB uses multiple iterative gradient boosters to construct a strong classification system [33,34] that has shown a preferable ability to classify imbalanced data in recent studies [32,35]. Due to the complexity of the final ensemble model, XGB can provide deep insight into the performance and has greater predictive power than most conventional methods [32]. A boosting strategy has been shown to be effective in controlling performance uncertainty [36]. Despite being widely used in many fields [31], XGB has also been effectively used in land cover classification [37–39], but rarely for imbalanced learning of remote sensing images.

To intuitively analyse the behaviours of the model with imbalanced data, a range of accuracy assessment techniques have emerged. Regular overall accuracy is used for overall agreement evaluation, which is the most widely used component of accuracy assessment in remote sensing [3]. However, overall accuracy greatly favours the majority class [12], leading to suboptimal classification and misleading conclusions [10,40,41]. Despite of the commonly used evaluation indicators for individual

class (user and producer accuracy) in the field of remote sensing, F-score which is encouraged in imbalanced learning is also used for quantifying the recognition efficiency on the minority class [10,12]. In addition to the evaluation of the agreement component, this paper also introduces disagreement components [42,43] to focus on the errors in the results. Moreover, in ensemble learning, the margin values represent the proximity of instances to the decision boundaries, and can be used to estimate the uncertainty of predictions [36]. Margin statistics can also reflect the effects of training data characteristics on classification outcomes [1]. Thus, this margin-based uncertainty measure is used to supplement to the traditional accuracy assessment methods to analyze the uncertainty of XGB for imbalanced remote sensing datasets with different training data characteristics (imbalance and spectral separability of the minority class).

Therefore, the goals of the present study are (1) to develop an XGB model for regional VHR image LCC with imbalanced training data, (2) to examine the overall performance, minority recognition and performance uncertainty of XGB on datasets with different class imbalances and spectral separabilities, and (3) to investigate the XGB's efficiency by comparing its performance with the well-known RF model. To evaluate the uncertainty, a margin criterion based on output probabilities is extended. Moreover, eight study areas with different classification taxonomies and complexity are used to test how the sample proportion and spectral separability of a specific class affect the XGB model under imbalanced conditions. After the introduction of the data (Section 2.2) and methods (Sections 2.2 and 2.3), the behaviour of XGB under different sample imbalances and spectral separabilities is described in Section 3.

2. Data and Methods

2.1. Study Areas and Data

The complexity of the study area can influence the imbalance of the dataset. Thus, the experiments are conducted in 8 study areas (1000 × 1000 pixels) with different complexities. Area 1 to 7 is located in Beihai, Guangxi, which is a typical suburb in a rural setting in China (Figure 1a). Area 8 is located in northwest Hobart, Tasmania, Australia (Figure 1b). The landscape distributions of the two regions exhibit different class imbalances without severe shadow occlusion from high buildings.

The experimental data are all VHR images. For area 1 to 7, the initial datasets are orthographic images with a spatial resolution of 0.2 m, acquired in 2014 by a Leica ADS40 digital camera. Three spectral bands are available: R (610–660 nm), G (535–585 nm), and B (430–490 nm). The dataset of area 8 is an open-access data source with a 0.5 m spatial resolution acquired in 2009 by Geo-Eye 1 (http://www.harrisgeospatial.com/docs/FXRuleBasedTutorial.html). VHR images can help manual interpretation obtain land cover types with high confidence, which is beneficial for random sampling and accuracy evaluating [44].

The species of area 1 to 7 are classified based on the China National Standards "Current Land Use Classification, GB/T 21010-2017" (http://std.samr.gov.cn/gb/gbQuery_) and ancillary data (the production of Second National Land Survey for Beihai). Meanwhile, to be consistent with the actual situation, the farmland class is further detailed into farmland with and without crops by manual interpretation (Figure 1a and Table 1). The land covers of area 8 are also referring GB/T 21010-2017 through manually interpreting. But the difference is that the building class is divided into three subcategories including Building-1 with red roof, Building-2 with grey roof and Building-1 with light green roof in pseudo mode, since the buildings have obvious "same objects with different spectrum" phenomena (Figure 1b in pseudo mode R/G/B:4/3/2). Additionally, high-light objects, such as ground vehicles, surface vessels, etc., are classified as one type. The species of area 8 are finally divided into 11 classes (Figure 1b and Table 1).

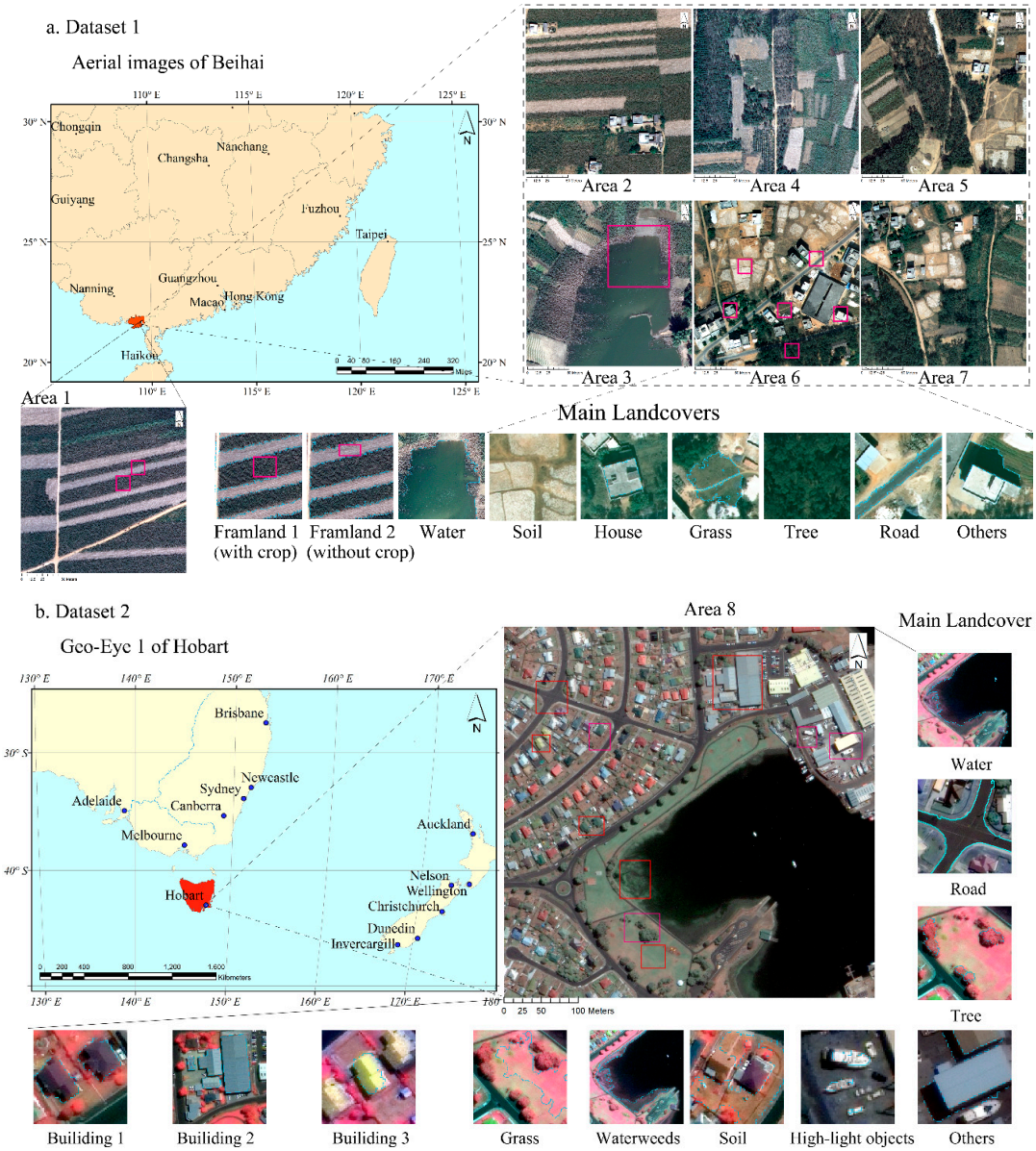

Figure 1. Eight study areas and the main existing land cover types (all classes in Beihai are in true colour mo de and classes in Hobart are in pseudo mode(R/G/B:4/3/2), expect 3 classes (road, highlight objects and shade). Dataset 1 of aerial images of Beihai is shown in a, and b is for dataset 2 of Geo-Eye 1 at Hobart.

Table 1. Shannon Diversity Index (SHDI) and species of the 8 study areas. (The boldface in the table is the minority class in the corresponding experiment.).

Area	SHDI	Dataset	Resolution	Species
1	0.83	ADS 40	0.2 m	Framland 1, Framland 2, **Soil**
2	0.94	ADS 40	0.2 m	**House**, Tree, Framland 1, Framland 2, Others
3	1.02	ADS 40	0.2 m	**Tree**, Framland 1, Framland 2, Soil, Water, Others
4	1.19	ADS 40	0.2 m	Tree, Framland 1, Framland 2, **Soil**, Grass
5	1.21	ADS 40	0.2 m	**House**, Tree, Framland 1, Framland 2, Soil, Others
6	1.43	ADS 40	0.2 m	**House**, Tree, Soil, Road, Grass, Others
7	1.67	ADS 40	0.2 m	**House**, Tree, Framland 1, Framland 2, Soil, Grass, Others
8	2.22	Geo-Eye 1	0.5 m	**Water**, Road, **Tree**, Builiding 1,Builiding 2, Building 3, Grass, Waterweeds, High-light Objects, Soil, Others

To quantify the complexity and diversity of the study areas, Shannon Diversity Index (SHDI), a common measure of species diversity that considers both species richness and abundance [45], are used. In order to conduct random sampling, evaluate the performance and calculate the area complexity, we make a reference data using the criteria above and manual interpretation. Thus, with this reference data, SHDI (H') can be calculated using:

$$H' = -\sum p_c \ln(p_c) \quad (1)$$

where p_c is the proportional abundance of class c relative to the total area of all species in a single study area. The SHDI of the 8 study areas ranges from 0.83 to 2.22 (acquired by Fragstats 4.2 [46]), meaning the diversity ranges from relatively simple to complex.

The digital number (DN) is used directly for feature extraction and classification. Although DN does not represent physical meaning, it can capture the differences among classes. As long as the training samples are derived from the images to be classified, the DN values can be applied to single date image classification [47]. In addition, 8 second-order texture metrics (mean, variance, homogeneity, contrast, dissimilarity, entropy, second moment and correlation) [48] for each spectral band are extracted in ENVI software with a 3 × 3 pixel template along the horizontal direction. All features are rescaled into the range [0,1] [49].

From the spectrum histogram of the two regions (Figure 2), we find that almost all the land cover types appear at certain intervals; in other words, overlapping among classes prevails in the spectral space, except for the water class in area 8, whose histogram is focused on a narrow interplot. To quantify the statistical distances between classes spectral distributions, we use well-known Jeffries-Matusita (JM) [50], which ranges 0 to 2.0. The closer that the value is to 2.0, the better the spectral separability is. From Table 2, the JM distances between water class and the others are almost perfect (2.0), while those of tree class have values around 1.8 (1.84 for tree and building-1, 1.82 for tree and grass). It is indicated that the spectral separabilities of water class and tree class have differences. Based on the difference, we analyzed the influence of spectral separability on imbalanced learning.

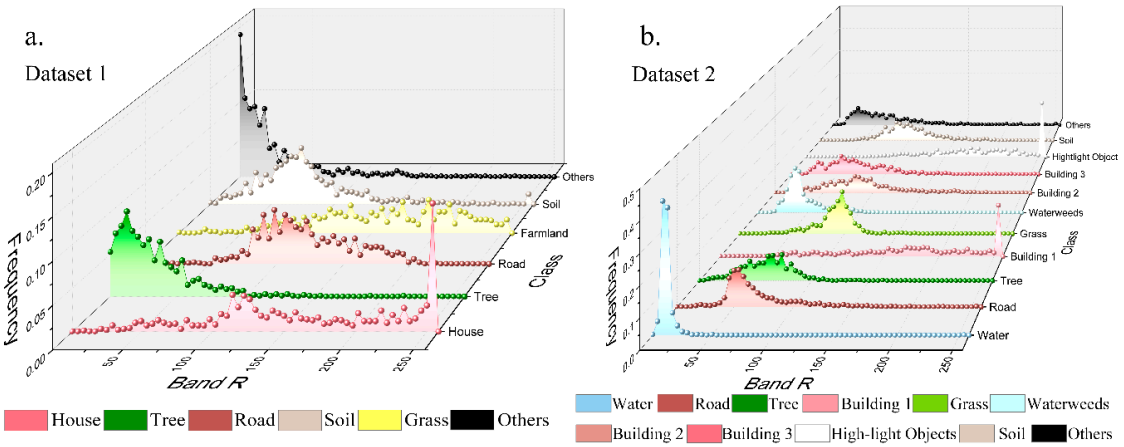

Figure 2. Spectrum histogram of land cover types in Beihai (**a**): R band and Hobart (**b**): R band.

Table 2. JM distance between Tree/Water class and the other classes in Area 8 (calculated using 1000 pixels per class with ENVI 5.1).

Pair Separation	Water	Road	Tree	Building 1	Grass	Water-Weeds	Building 2	Building 3	Highlight Objects	Soil	Others
Tree	2.00	2.00	-	1.84	1.82	1.98	2.00	2.00	2.00	2.00	1.97
Water	-	2.00	2.00	2.00	2.00	1.99	2.00	2.00	2.00	2.00	1.98

2.2. Methods

The methodology to explore the efficiency of XGB with imbalanced data was to generate random sets of training samples with distinct proportions between the minority and majority classes. The complete workflow, model construction (Section 2.2) and performance evaluation (Section 2.3) is depicted in Figure 3. To analysis the efficiency of XGB, the performance is compared with RF, which is also used for imbalanced learning for remote sensing data [1], under the same datasets.

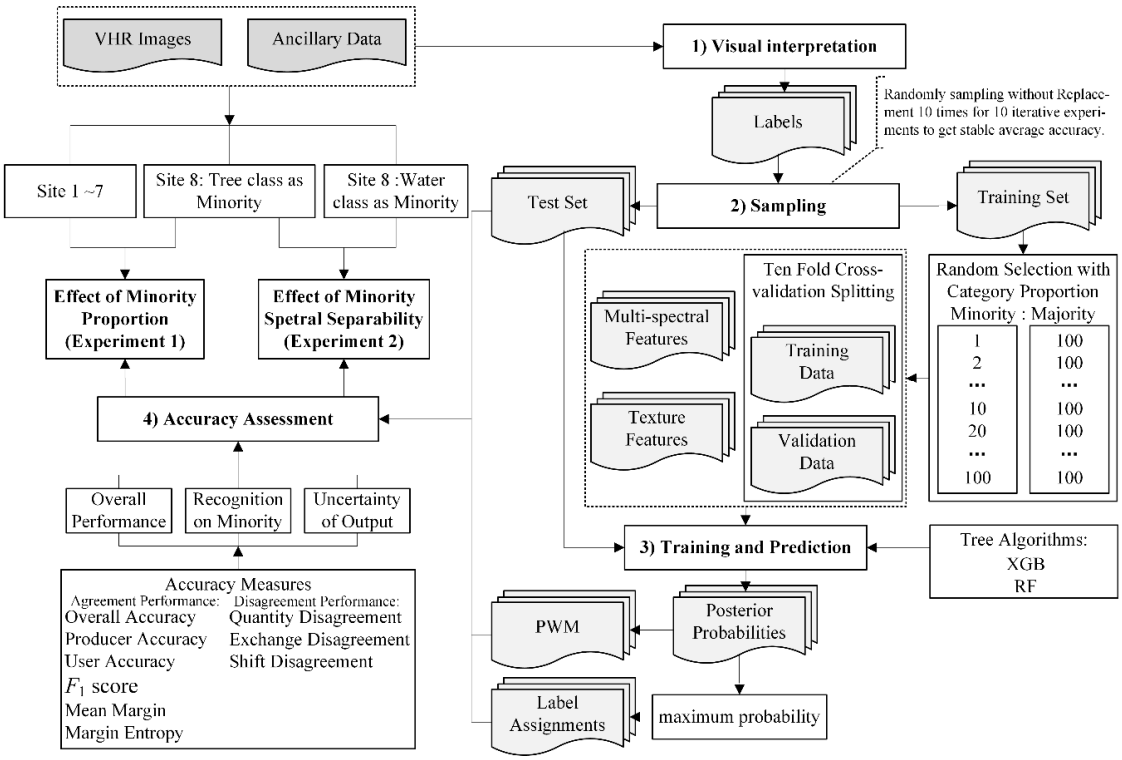

Figure 3. Overview of the workflow for extreme gradient boosting (XGB) with imbalanced data.

2.2.1. Models and Parameters Optimization

Extreme gradient boosting (XGB) [30] is a tree-based ensemble learner that uses regression decision trees (DTs) as baseline booster [51]. Due to the regularization, gradient descent boosting in XGB is more predictive than standard boosting techniques [32]. Comparing with other gradient boosting trees, XGB can be implemented concurrently and takes measures to reduce the risk of over-fitting [30]. This makes it very attractive in imbalanced issues. XGB was successfully applied in a land cover classification scheme [38] and imbalance learning [52,53]. Here is a simple methodology presentation of XGB.

Suppose, for a given data set with n examples $\mathcal{D} = \{(x_i, y_i)\}, (i = 1, 2, \ldots, n)$, y is the label of instance x, and a booster DT can be denoted as $T(x)$. Moreover, there are M boosters in the ensemble system, denoted as:

$$f_i = \sum_{m=0}^{M} T_m(x_i) \tag{2}$$

When $f_i = y$, it means the correct prediction is given, otherwise, wrong label was given. For m-th tree training, i.e., the m-th loop for the system constructing, there are two main steps. Firstly, computes the current pseudo-residual based on the output of the $(m-1)$-th loop, and then use the

attribute-residual set to train the m-th sub-tree. In XGB, additional regular item (Equation (3)) is added in the loss function (Equation (4)).

$$\Omega(T_i) = \gamma \cdot Leaf + \frac{1}{2}\lambda\|\omega\|^2 \tag{3}$$

$$\Theta = \text{argmin}\sum_{i=1} Loss(f_i, y_i) + \sum_m \Omega(T_m) \tag{4}$$

where, leaf in Equation (3) is the tree complexity penalty coefficient, and ω is the magnitude of leaf weights. Thus, the objective function to minimize in the m-th round is:

$$\Theta^m = \sum_{i=1}^{n} Loss\left(f_i^{m-1}, y_i + T_m(x_i)\right) + \Omega\left(T_j\right) \tag{5}$$

In XGB, objective functions are simplified and approximating with Taylor expansion and L2 norm. Subsampling, randomly partition a subset of the training samples used for multiple training, can be adopted in XGB for controlling overfitting. Sparse aware is used against missing values, and built-in cross validation allows users to get the optimal iterations amount in a single run. Column subsampling reduces computational complexity. In addition, XGB is parallel at feature level, and takes an additional second term function as regularization to smooth the final learnt weights which is not under account in similar models. It is a model which can draw on superiority both of statistical and ML techniques. Due to all these advantages, it may be suitable for imbalanced remote sensing data land cover classification.

According to the principle above, appropriate parameter combination is necessary to ensure that the model is well generalized and predictive. Therefore, we employed the grid search and ten-fold cross-validation to perform parameter optimization for XGB [32]. Five main parameters are optimized, in detail: n_estimators (number of boosted trees to fit), learning_rate (step size shrinkage used in update to prevent overfitting), max_depth (maximum depth of a tree), min_child_weight (minimum sum of instance weight needed in a child), and subsample (subsample ratio of the training instances) (XGBoost Tutorials, https://xgboost.readthedocs.io/en/latest/tutorials/index.html). First, we used cross-validation (with an initial value set to 1000) to find a rough value for n_estimators. Then, under this n_estimators value, we used grid search in three steps to find more appropriate values for learning_rate (the lower and upper bounds for search are 0.1 and 1), max_depth (the lower and upper bounds are 2 and 23) and min_child_weight (the lower and upper bounds are 1 and 29), subsample (the lower and upper bounds are 0.5 and 1). Then the best values are searched near the found result values above with smaller steps. During the process, with the changing values of other parameters, n_estimator and learning_rate are reciprocally optimized. Parameters for all study areas are optimized using the same producer. Taking area 6 as an example, the best parameter values were found to be 7 for max_depth and 8 for min_child_weight, 99 for max_depth, 0.09 for learning_rate, and 0.8 for subsample.

As a typical remote sensing classification model, RF is also used in the classification with imbalanced remote sensing sample sets [1]. Thus, we have the performance of XGB compared with RF under the same datasets. Comparing with gradient boosting approach, the trees in RF are built independently to each other, while in XGB, new tree always trains based on the results of the previous round. According to the previous studies [54,55], two main parameters were set for RF: the number of the decision trees to form the forest model (n_estimators) and the size of the random subsets of features to consider when splitting a node (max_features). Theoretical and empirical research has proven that the accuracy performance of RF is insensitive to the n_estimators parameter as long as the total number of subtrees is large enough [54]. To ensure that the complexities of the models are in the same order of magnitude, n_estimators for RF is equal to the optimization parameters of XGB (n_estimators). The max_features parameter is set to the square root of the number of predictor variables [1,54].

2.2.2. Experiments: Analysis of Class Imbalance and Spectral Separability

Two experiments were conducted to analyze the effects of skewed data distributions and spectral separability on XGB. To analyse the influence of the minority proportion (Experiment 1), classes of low proportion in area 1 to 7 (bold text in Table 1) and the tree class in area 8 were assigned to be the minority classes. For Experiment 2, the tree and water classes in area 8 were assigned as the minority class, presenting a striking contrast in spectral separability that affects remote sensing imbalanced learning. (Figure 3)

The main concept behind this exploration is to generate random sample sets with different imbalance proportions and analyze the performance of models on these sample sets iteratively. Each classification iteration is based on the training sets consisting of the same patterns. For multiclass training sets, we manually assigned one class as the minority class and the others as majority classes using equal sample amounts (1000 per class) in the training set. Then, with random selection, the minority proportion changes from 1%, 2%, . . . , 10%, 20%, . . . , 100% (balanced) of the amount for the majority, forming distributions ranging from extremely skewed to completely balanced. We conducted ten reproducible trials with independent randomly selected training sets of each individual iteration to ensure the stable accuracy. The final reported results of each accuracy measure are the average values of the ten trials.

2.3. Accuracy Assessment

Considering the skewed distribution of the limited reference dataset in remote sensing classification situations, both the global accuracy tendency and the partial variation in the minority class are observed. Therefore, we use classical confusion matrix (CM)-based accuracy metrics to obtain the overall performance and recognition efficiency for the minority class. In particular, the extended probability-weighted margin (PWM) descriptive statistics [1] and these margin-based measures are used to evaluate the uncertainty for both instances and models, respectively.

2.3.1. CM Based Accuracy Metrics and Disagreement Performance

For overall performance comparison, overall accuracy (OA) and Cohen's Kappa coefficient (κ) are the most widely used measures in remote sensing [3,50]. However, Pontius (2011) [42] indicated that the κ is useless and misleading from a practical perspective because of its flawed methodology. The authors recommend a simpler and more appropriate measure focusing on the errors, which we call the disagreement performance in this paper, including three components: quantity [quantity disagreement (QD)], exchange [exchange disagreement (ED)] and shift [shift disagreement (SD)] [43]. Three components can be simply calculated using CM, and details can be found in [43]. Thus, OA and overall disagreement performance are adopted to access the precision and errors of the overall performance in this paper.

In remote sensing, producer and user accuracies are widely used for evaluating specific classes. Thus, they are used for quantification of the performance of a specific class. However, in this imbalance machine learning scenario, the ideal classification result of a specific class is for the precision and recall to be both relatively stable and high. The F-score [12] metric considers both precision and recall at the same time. Thus, we use the balanced F-score ($F1$), calculated as $F_1 = 2precision \times recall/(precision + recall)$, to assess the recognition efficiency of the minority class. Similarly, the disagreement performance of a specific class [43] is used for analyzing the nature of errors. The calculation details can be found in [43] and Appendix A.

In summary, we dissect not only precision for both overall and specific classes but also the nature of errors to understand the influence of the imbalanced data on the classifiers' performance (Figure 3).

2.3.2. The Extended Margin and Margin Based Confidence Measures

The ensemble margin is an important factor in ensemble learning [1,36]. Most ensemble systems use a voting classification rule to obtain the final hypothesis for an instance. Therefore, a classification margin was proposed to measure the proximity between the instances and the decision boundary under this voting mechanism. The margin is typically simply defined as the difference between the number of correct votes and the maximum of the incorrect votes. Obviously, an instance is correctly classified if and only if the ensemble margin is positive; thus, a larger positive margin indicates a correct prediction with lower uncertainty [36]. However, to thoroughly analyze the generalization error of a model, it is necessary to consider not only the accuracy of the model prediction but also the uncertainty of the output [36].

Margin functions can be defined using various formulas [56]. Mellor [1] provided new margin statistics for RF with regard to the remote sensing imbalanced issue. However, the margin criterion calculated by Mellor was based on the voting results by all the weak learners in the ensemble system. In this paper, we use the output probability instead of the votes in RF to extend Mellor's margin criterion in order to make this margin measure suitable for non-ensemble classifiers and non-voting strategies. The PWM is defined as the difference between the largest and the second largest output probability for an instance, meaning the difference between the most and the second most potential classes:

$$PWM(x) = max_{i \in C} P(c_i|x) - max_{j \in C, i \neq j} P\left(c_j \middle| x\right) \quad (6)$$

where $P(c|x)$ is the probability of instance (x) to be predicted as class c and C is the class set. PWM ranges from 0 to 1, where 1 denotes a prediction with the highest confidence level; while if it is a binary classification, 0 implies that the result is no better than a random guess. In other words, the PWM reflects the confidence of the prediction to some extent. Of course, the correct classification is expected to have high PWM and relatively low value for misclassification. Indeed, lower the PWM of a misclassification, more opportunity to improve the performance [1].

We can also calculate the confidence measures (mean margin and margin entropy) proposed by [1] using this PWM, instead of using voting-based margin. Hence, in this paper, we use this PWM based mean margin and margin entropy to estimate the performance confidence, and PWM for the confidence of predictions for individual instances. Mean margin is defined as the average value of the difference between the margin associated with correctly classified and misclassified instances. A higher mean margin value means that the classifier prediction involved higher probability and is more correct. The margin entropy measures the diversity and redundancy of all the PWM for single instances using Shannon's entropy. We also used 10-bin histogram as Mellor did for this margin entropy calculation. The entropy is expected to be high but not reach the maximum (for the 10-bin histogram, the maximum of entropy is 3.32 when the frequency of all bins is equal). Whilst a large output diversity is unnecessary for achieving a high-confidence performance [57], it is also indispensable for ensemble learning systems. Specifically, when the output of all the baseline learners is homogeneous (less diversity and low entropy), the ensemble system has no superiority over baselines, whereas output that is too varied (more diversity and high entropy) is tantamount to random guesses. The relationship between diversity and ensemble accuracy is quite complex [57]. Therefore, an appropriate margin entropy value should be high but not reach the maximum [1]. The calculation details can be found in [1] and Appendix A.

The class probabilities of XGB and RF can be calculated by Application Program Interface (API). For XGB, the predicted probability is the summation of the bias (calculated by the boosters in the first training round) and the outputs in each boosting round. It acquires the class probabilities by setting the parameter objective as multi:softprob (https://xgboost.readthedocs.io/en/latest/parameter.html). For RF, the class probabilities are the average probabilistic predictions of classifiers in RF, which can be calculated by scikit-learn package (https://scikit-learn.org/stable/user_guide.html). The PWM and the confidence measures of XGB and RF are finally computed using the class probabilities.

3. Results

To analyze the influence of the sample imbalance in these multi-class classification scenarios, all the results are described in 3 aspects: overall performance, recognition efficiency on the minority class and performance uncertainty. For all the performance of overall and specific class, not only the correct predictions are considered but also the errors.

3.1. Experiment 1: Minority Proportion Influence on XGB

1. Overall performance. The overall accuracy of XGB for all 8 areas across different sample imbalances is shown in Figure 4. Among all the experiments, the highest OA values appear when the training dataset is either balanced or almost balanced, when the minority proportion is 90% of the majority, which is consistent with the results in [1]. From Figure 4, we can see that the OA improves steadily as the minority class proportion increases. When the ratio of the minority to majority classes is 50:100, XGB achieves approximately 90% of the whole development of all 8 areas. In fact, even at minority:majority proportion of 30:100, approximately 80% of the benefits can be achieved except at area 3 (58.67%). In a particularly simple scene, such as area 1, even the sample set is extremely imbalanced (minority:majority < 10:100), and the OA is significantly improved (88.12%). The results demonstrate that the improve speed is not always uniform across the change of sample imbalance. A small increase in sample for the minority class can lead to a considerable increase in overall accuracy when suffering extremely imbalanced. However, a trade-off can be found between overall accuracy and sample requirement. The OA of XGB for area 8 (which improved by only 7.1%) is more stable than its performance for area 6 (improved by 12.5%) across all the minority class proportion increases. Since area 8 includes more species, the training set for area 8 suffers much more from imbalance than that of area 6. Thus, the extreme imbalance limits the contribution of the minority class to the overall accuracy in the complex scenario. Apparently, the trade-off is related to the complexity of the study area and the contribution of the minority to overall accuracy. Comparing with RF (red line in Figure 4), with the increase of the complexity of the study area, the performance of XGB in terms of overall accuracy is higher than that of RF under the same experimental conditions.

To understand the errors, we analysed the overall disagreement performance of all classes. The disagreement performance (QD, ED, and SD) across different sample imbalances is shown in Figure 5. It is obvious that the overall QD changes significantly with extremely imbalanced and balanced samples, even though the general error (accumulation of QD, ED, and SD) becomes relatively stable after the sample imbalance achieves minority:majority ≥ 50:100. However, ED and SD among different sample imbalances are not as variant as QD, especially SD. It is interesting that, as the sample imbalance decreases, the ED slightly increases. This pattern indicated that sample imbalance mainly influences the overall quantity disagreement and that balanced samples do not ensure the best overall exchange disagreement. Comparing with RF (red bars in Figure 5), with respect to error components, the sample imbalance mainly influences QD for both XGB and RF. In general, XGB can provide comparable overall performance to RF when suffering imbalanced samples, even slightly higher in more complex scenarios, since providing sight advantages in error components.

2. Recognition efficiency of the minority class. We take the visual maps and result values of area 6 and area 8 for detailed description, and all areas for variation tendency diagnosis. The visual classification maps of XGB are shown in Figure 6 (for area 6) and Figure A1 (for area 8).

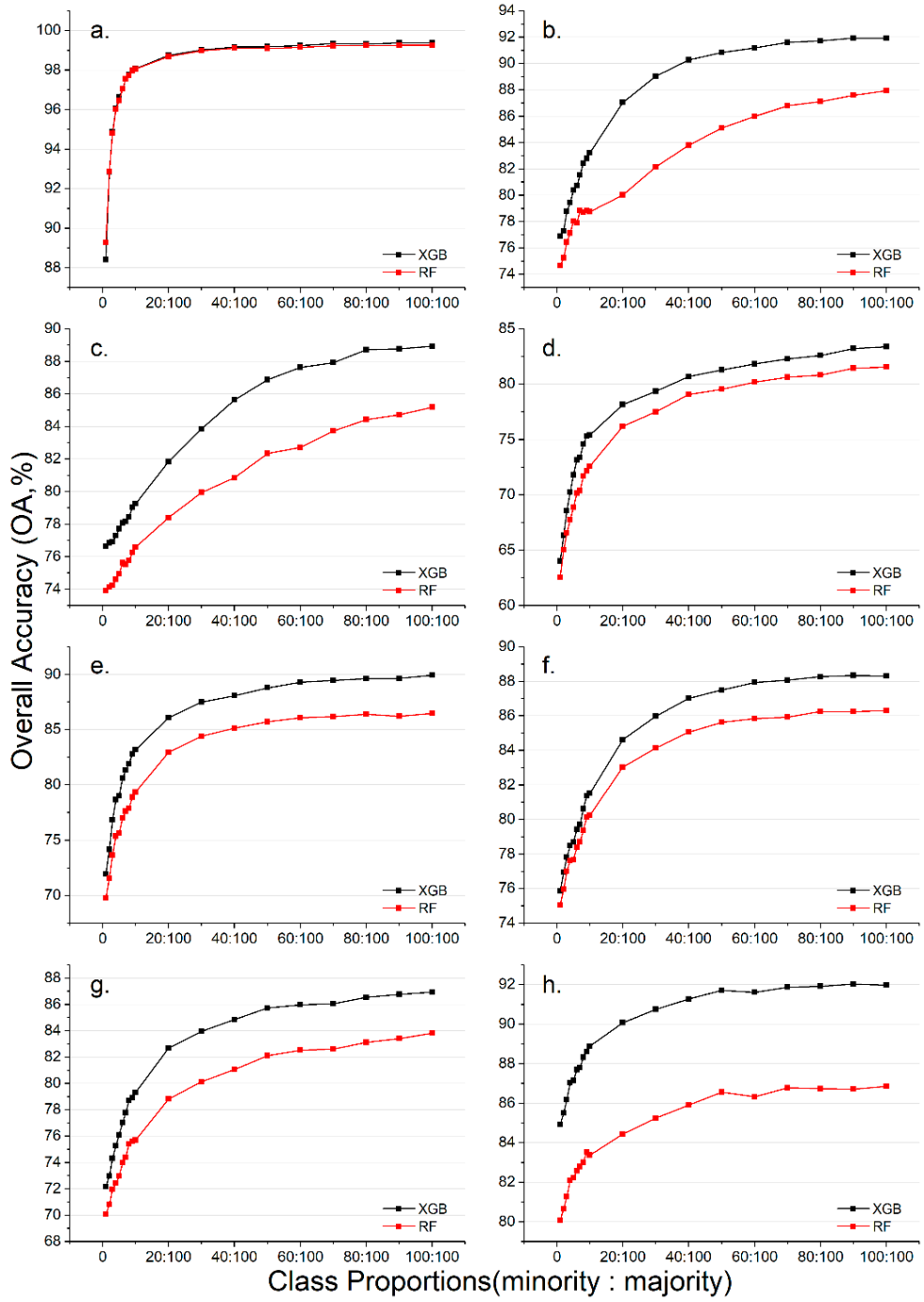

Figure 4. Overall accuracy curves of XGB (black) and random forest (RF) (red) for 8 areas across different minority proportions (minority: majority from 1:100 to 100:100) of the training data set. The results of different areas are at: (**a**) area 1, (**b**) area 2, (**c**) area 3, (**d**) area 4, (**e**) area 5, (**f**) area 6, (**g**) area 7, (**h**) area 8.

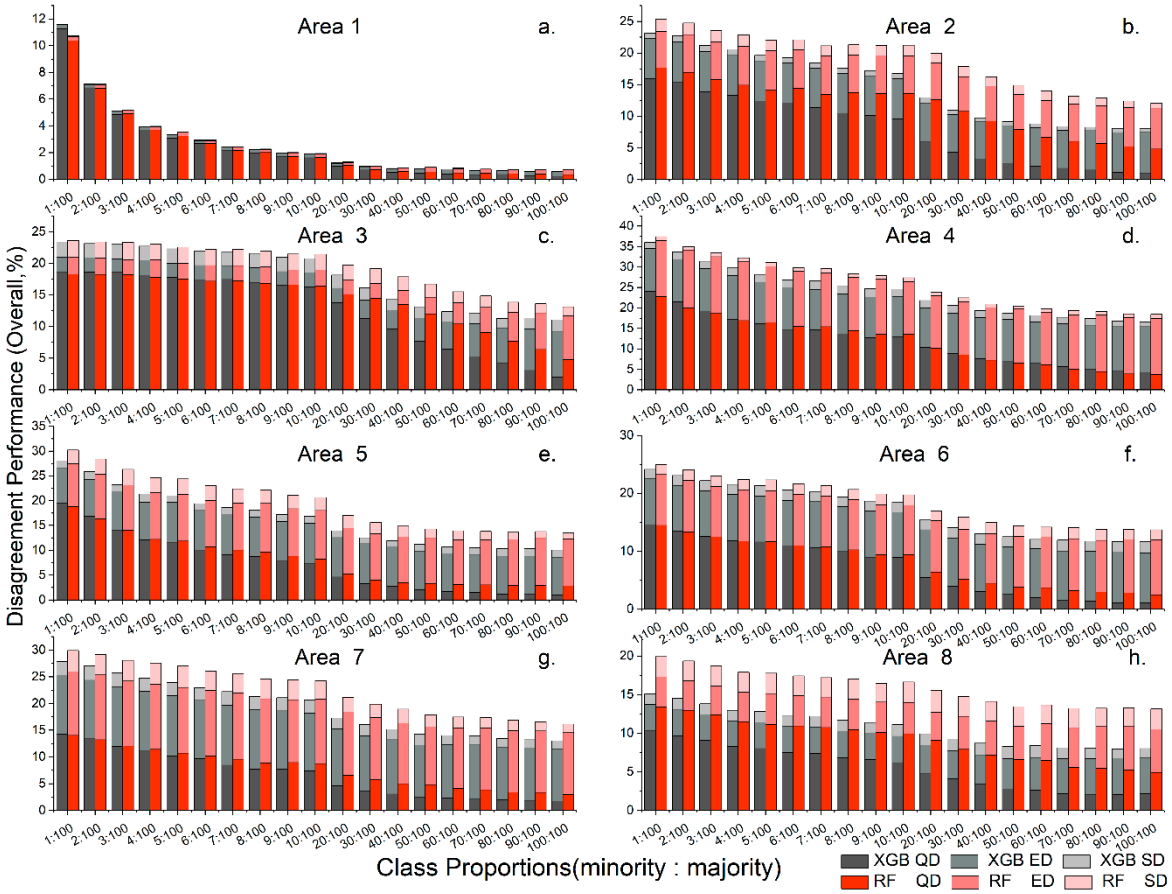

Figure 5. Stacked bars of overall disagreement performance of XGB (grey) and RF (red) for all 8 area across different sample imbalance. The results of different areas are at: (**a**) area 1, (**b**) area 2, (**c**) area 3, (**d**) area 4, (**e**) area 5, (**f**) area 6, (**g**) area 7, (**h**) area 8.

Obviously, the identification of the house class (the minority class) for area 6 varies greatly as the training data distribution changes. When the training data have an extremely skewed distribution (minority:majority = 1:100, Figure 6b), almost all the houses are misclassified as roads, such as, the parts marked with the blue frames in Figure 6. In fact, only 38 of all the 158,292 pixels representing the house class are identified correctly. When the minority:majority ratio increases to 10:100, the classifier correctly classifies 42,971 of the targets, nearly 1131 times the number classified correctly at 1:100. When the ratio reaches 50:100, the number of correct classifications of the minority class is 89,971, which is approximately 57% of the entire house class. Finally, a correct recognition rate of 76% is achieved when samples are balanced between classes.

However, the producer and user accuracies of the minority class are affected by the lack of samples, and user accuracy is mainly affected when the imbalance ratio (minority:majority) is larger than 10:100. Figure 7a shows the performance on producer and user accuracies of the minority class (House) and one of the majority class (Soil). When imbalance ratio is under minority:majority = 5:100, both the producer and user accuracies are quite low. After this ratio, the producer accuracies for the minority class are almost 1, but the user accuracies are under 0.7 until the ratio reaches 40:100. That means, the minority class instances identified by the classifier are almost all correct, but the omission rate is actually high. Meanwhile, the same accuracy metrics of the majority are less affected by sample imbalance. Take soil class as an example (Figure 7b), the producer accuracy changes within 4%, while user accuracy changes within 10%. Sample imbalance has a great impact on user accuracy under imbalance learning scenario. However, the user accuracy of the XGB is higher than that of RF under the same experimental conditions, with a maximum advantage above 11% (Figure 7a).

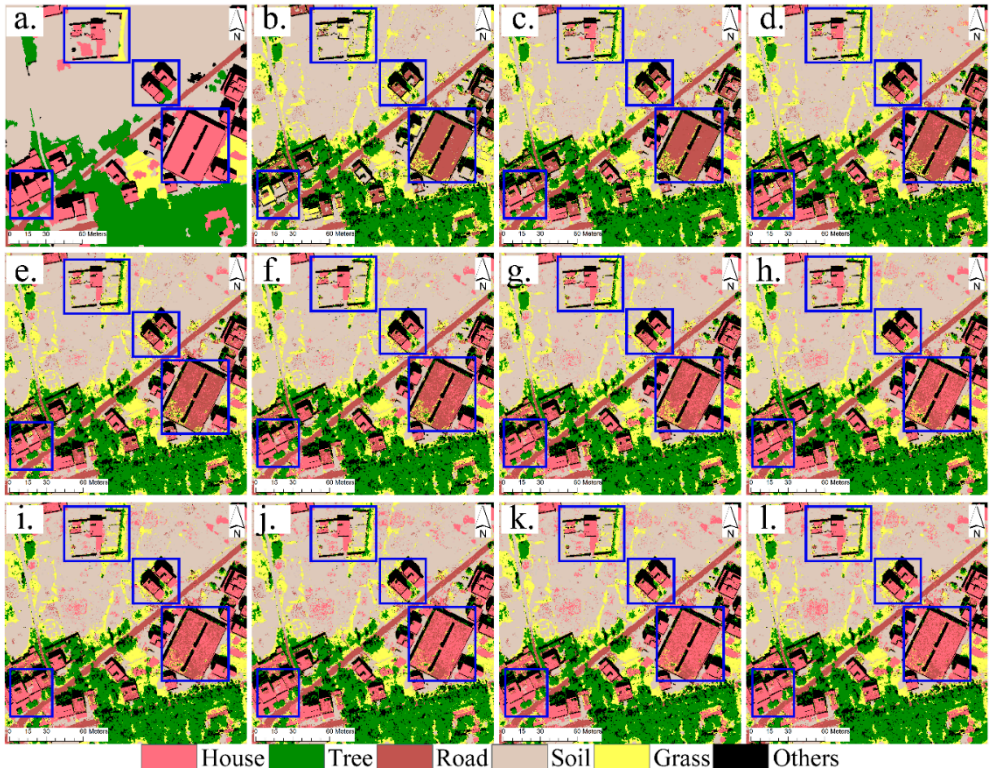

Figure 6. Reference map (**a**) and Classification maps of XGB using the training sets with the ratio of minority: majority at: (**b**) 1:100, (**c**) 10:100, (**d**) 20:100, (**e**) 30:100, (**f**) 40:100, (**g**) 50:100, (**h**) 60:100, (**i**) 70:100, (**j**) 80:100, (**k**) 90:100, (**l**) 100:100, for area 6.

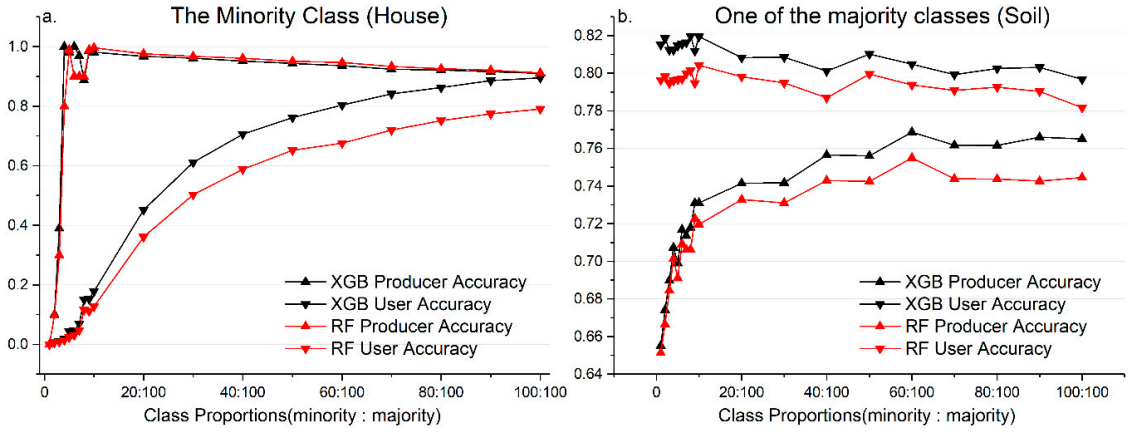

Figure 7. Curves of producer and user accuracies for the minority (**a**: house) and one (**b**: soil) of the 5 majority classes for area 6 across different sample imbalance.

In fact, compared to producer and user accuracy, the *F*1 score is a more intuitive indicator to describe the the recognition efficiency of the minority class under different imbalances in the training data (Figure 8). The growth rate (scope of curve) of the *F*1 score curve shows different trends as the sample imbalance changes. For example, for area 6 (Figure 8f), when the sample imbalance ratio is between 1:00 and approximately 40:100, the improvement speed is significantly faster than that above 40:100 and not as fast as the speed in the interval of 1:100 to 20:100. Even if the amount of sample for the minority class is increased from 40% of the majority class to 100%, the improvement in *F*1 score is increased from approximately 0.8 to approximately 0.9, an improvement of only 0.1. Although

improving the *F*1 score to 0.8 in area 6 requires a sample imbalance ratio from 1:100 to 40:100, while area 8 requires a ratio of 50:100, the characteristics of the two curves are very similar. This similarity implies that the sample imbalance is the main impact factor under relative imbalance learning for the minority classes in area 6 (house class) and area 8 (tree class), and as the imbalance decreases, the dominant influence gradually decreases. However, for areas 1, 4 and 5, the coincident curves imply that the performances of XGB and RF are quite similar. For the other areas, XGB performs better than RF (its maximum advantages are area 2: 0.49; area 3:0.29; area 6: 0.09; area 7: 0.18; area 8: 0.23). The largest *F*1 scores of the two methods are quite close, with a maximum difference of 0.076 (area 7).

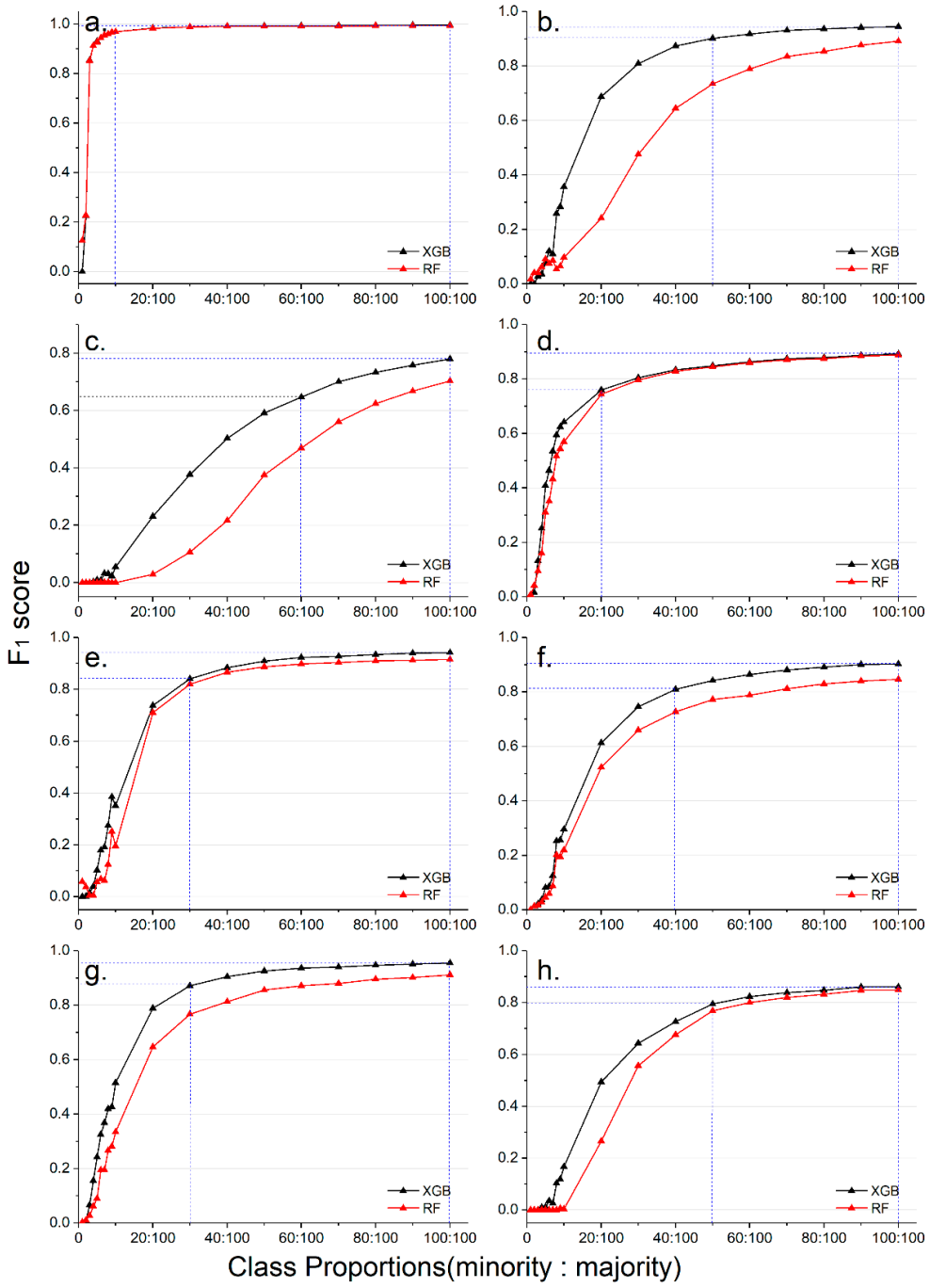

Figure 8. The F_1 score curves of XGB and RF for all areas across different minority proportions (minority: majority from 1:100 to 100:100). The results of different areas are at: (**a**) area 1, (**b**) area 2, (**c**) area 3, (**d**) area 4, (**e**) area 5, (**f**) area 6, (**g**) area 7, (**h**) area 8.

To capture the error components of the minority class, the disagreement performance of house class for area 6 is shown in Figure 9a, and in contrast, Figure 9b shows the same measures for one of the five majority classes, soil class. From Figure 5, the main error components of the overall disagreement is QD, and the minority class contributes almost 96.02% (14.59% divided by 14.01%) to 85.53% (2.53% divided by 2.16%) of the QD (Figure 9a), when the sample imbalance meets minority:majority < 50:100. As the sample proportion of the minority class increases, it brings a decrease in QD for both the house and soil classes (Figure 9b) but a slight increase in EQ and more mild changes in SQ. It can be inferred that adding samples of the minority class will not only change the error components for the minority class but also the majority class to various extents.

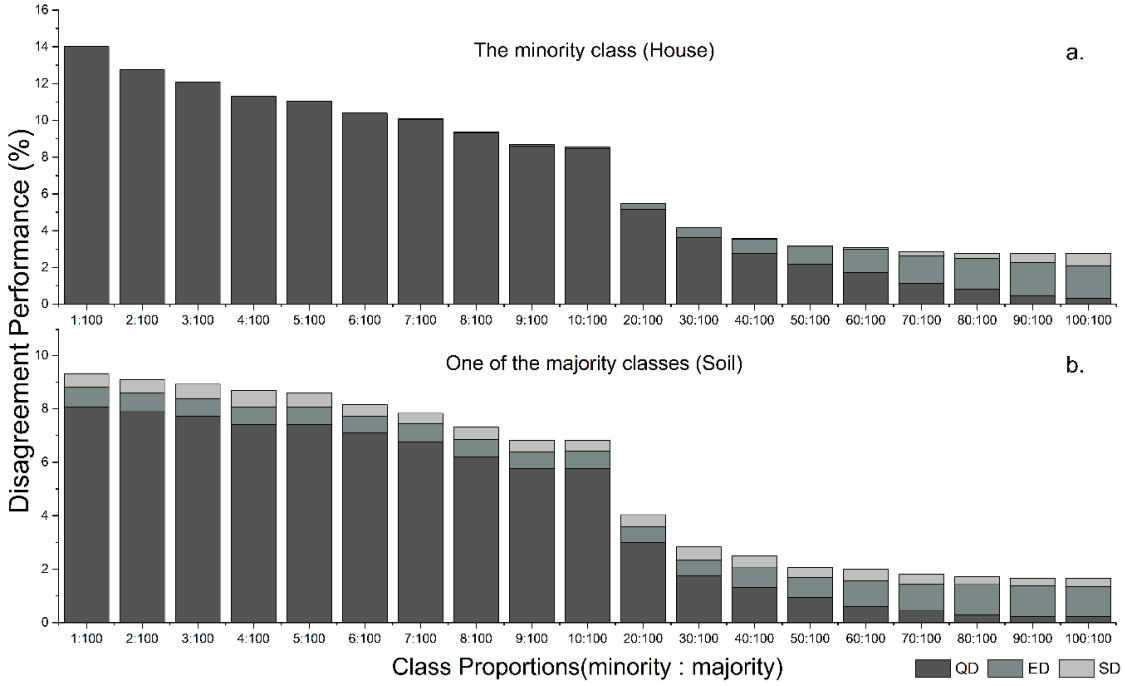

Figure 9. Stacked bars of disagreement performance for the minority (**a**: house) and one (**b**: soil) of the 5 majority classes for area 6 across different sample imbalance.

Comparing with RF, for all areas, the error components of the minority class for both XGB and RF are quite similar across different sample imbalances (Figure 10). In some experiments (areas 2–8), the error is slightly higher than in RF, while the difference is within 1%. The cumulative error displayed in each study area is consistent with the results reflected by the *F*1 score curves (Figure 8).

3. Performance uncertainty. Figure 11 shows the cumulative frequency distribution curves for the PWM of XGB when the training dataset has different class proportions (minority: majority = 1:100, 10:100, 50:100 and 100:100), grouped by the correctly classified and misclassified instances for area 6. XGB achieves performances with more than 90% of the high margins (>0.5) both for correctly and incorrectly classified instances with different sample imbalances (even with minority:majority = 1:100, the percent is 96.68%). With different imbalanced samples, the margin cumulative frequency curve of the correctly classified examples is very close, with a maximum difference within 8%. Meanwhile, the margin distribution of misclassified instances is obviously affected by the sample imbalance, with a maximum difference above 23%. As the proportion of samples for the majority class increases, the misclassified instances exhibit higher margins compared with the relatively imbalanced scenario. It can be seen that relatively imbalanced samples will not significantly reduce the confidence associated with correctly classified instances of XGB but affect the uncertainty of misclassification instances.

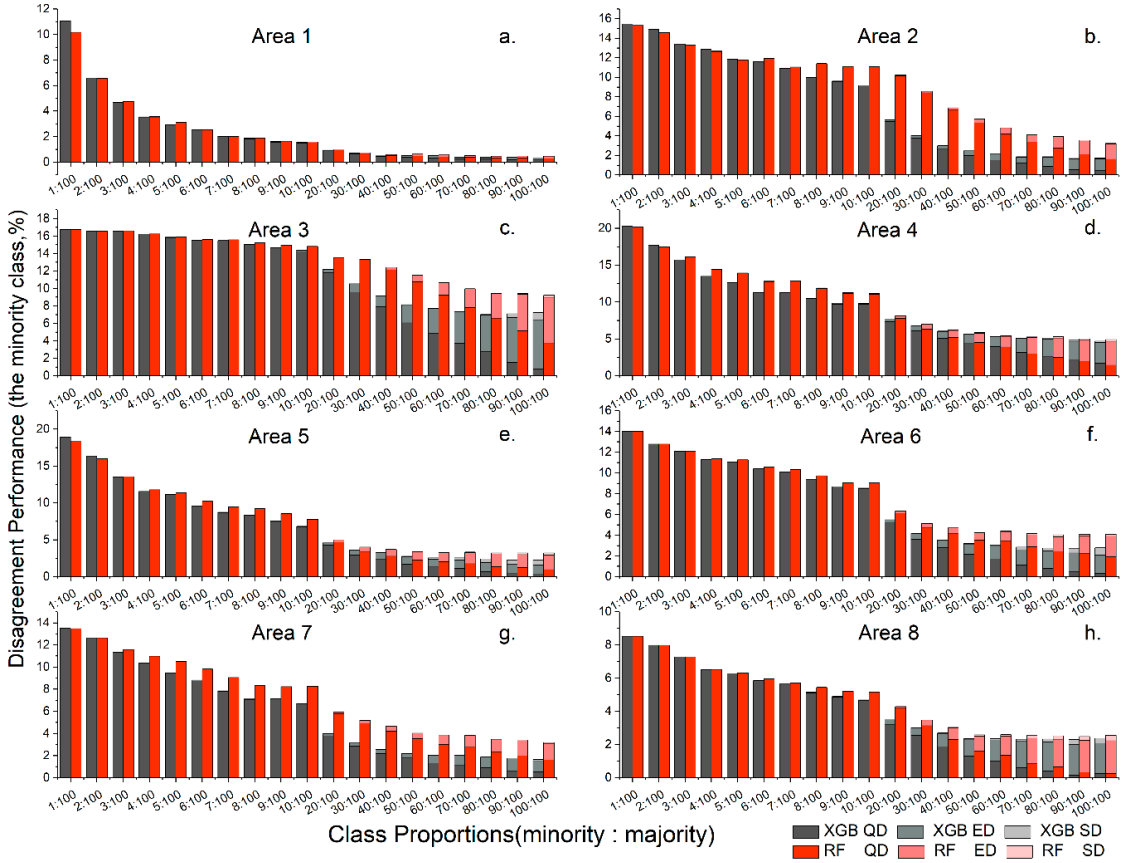

Figure 10. Disagreement performance of the minority classes using XGB (grey) and RF (red) for all areas. The results of different areas are at: (**a**) area 1, (**b**) area 2, (**c**) area 3, (**d**) area 4, (**e**) area 5, (**f**) area 6, (**g**) area 7, (**h**) area 8.

The margin-weighted confusion matrix demonstrates that the performance regarding per-class uncertainty is associated with correctly classified instances [1]. Taking area 6 as an example, Table 3 shows the CM and margin-weighted confusion matrix of the XGB and RF using the same training data, where the minority:majority ratio is 40:100. Compared with the recall from CM, the maximum difference between the corresponding values is 3.22% and the minimum difference is less than 0.2%. In terms of per-class precision, except for the minority class (house), which has a difference of 7%, the differences between other classes are below 1.5%. Thus, the CMs of the two models behave very similarly. However, the matrixs of XGB shows notable improvement over that of RF with respect to uncertainty associated with correctly classified instances (the cells on the main diagonal), with an average of 0.82, which is 46% greater than the corresponding values of RF. This finding implies that the correctly classified instances of XGB have a significantly lower uncertainty than those of RF. However, the non-main diagonal cells of average PWM are also much higher than those of RF, which are expected to be low. In terms of the PWM associated with the misclassification, RF receive better uncertainty. Considering the absolute amount, the improvement of certainty associated with correct classification is more dominant.

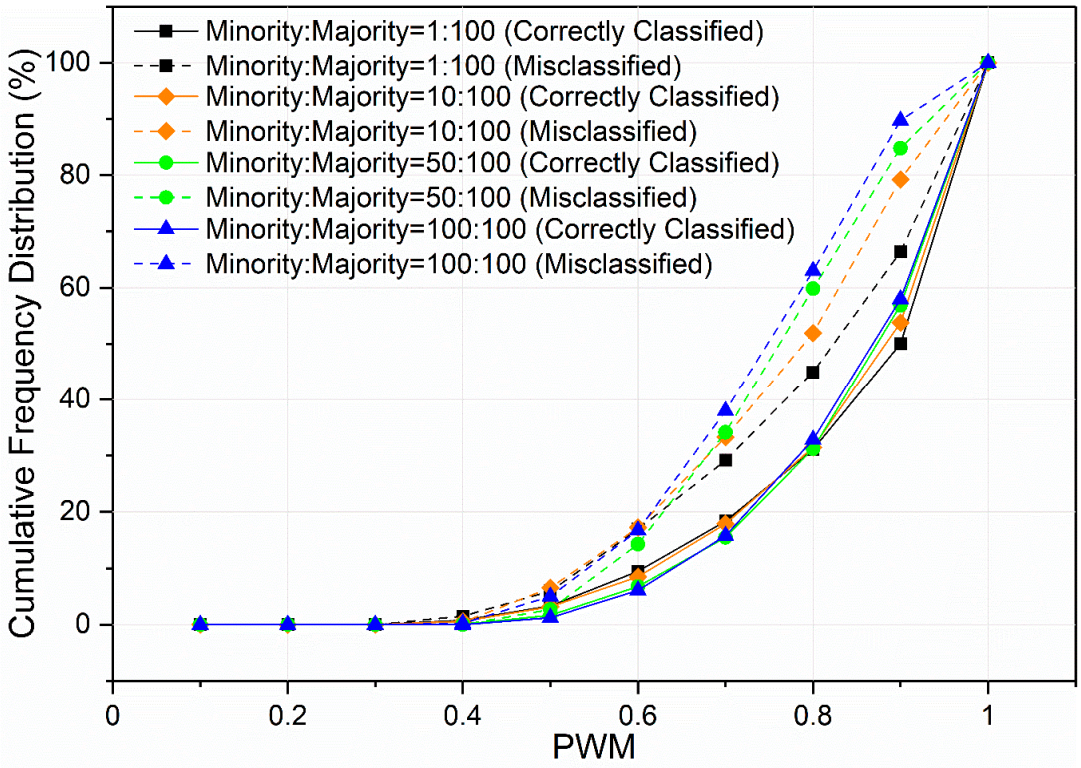

Figure 11. The probability-weighted margin (PWM) cumulative frequency distribution curves of area 6 grouped with correctly and misclassified instance of XGB using the training sets with the ratio of minority: majority at: 1:100 (dark), 10:100 (orange), 50:100 (green), 100:100 (blue).

Table 3. Confusion matrix (CM) and margin-weighted confusion matrix of XGB and RF of Area 6 with training set suffering a ratio of minority and majority at 40:100 (One of ten trails).

		Minority Proportion in Training Data Set: 40% of (per) Majority											
		CM						**Margin-Weighted Confusion Matrix**					
		House	**Tree**	**Soil**	**Road**	**Grass**	**Others**	**House**	**Tree**	**Soil**	**Road**	**Grass**	**Others**
XGB	House	293	8	33	33	1	11	0.80	0.67	0.86	0.78	0.58	0.84
	Tree	1	498	2	3	65	35	0.81	0.78	0.95	0.78	0.66	0.83
	Soil	1	0	636	9	1	1	0.68	0.00	0.92	0.84	0.77	0.84
	Road	3	1	4	589	56	2	0.76	0.82	0.94	0.86	0.69	0.89
	Grass	4	46	1	30	392	2	0.75	0.72	0.96	0.80	0.69	0.76
	Others	2	29	2	3	13	631	0.86	0.68	0.95	0.79	0.66	0.88
		House	**Tree**	**Soil**	**Road**	**Grass**	**Others**	**House**	**Tree**	**Soil**	**Road**	**Grass**	**Others**
RF	House	265	8	56	37	2	11	0.53	0.29	0.47	0.43	0.25	0.38
	Tree	1	494	2	3	66	38	0.41	0.48	0.76	0.46	0.35	0.66
	Soil	0	0	635	10	2	1	0.00	0.00	0.67	0.37	0.22	0.11
	Road	3	3	4	587	56	2	0.41	0.22	0.39	0.52	0.46	0.78
	Grass	4	51	2	31	385	2	0.39	0.44	0.45	0.42	0.45	0.69
	Others	2	37	2	3	12	624	0.52	0.37	0.64	0.39	0.32	0.72

Mean margins, when evaluating the uncertainty of the whole output, are expected to be as high as possible [1]. For all the tested areas, XGB has higher mean margins, and its maximum values are approximately twice those of the corresponding RF values or even higher (Figure 12). Moreover, comparing the distribution of the mean margin values of the 8 study areas, we observe that the mean margin values for the relatively complex area (area 6, 7, and 8) are more stable (the boxes are narrower)

than those of the remaining areas. However, the complexity of the training data aggravates the class imbalance. The performance uncertainty of XGB is stable and large in terms of mean margins (>0.5) for relatively complex imbalanced datasets (area 6, 7, and 8) and much higher than that of RF.

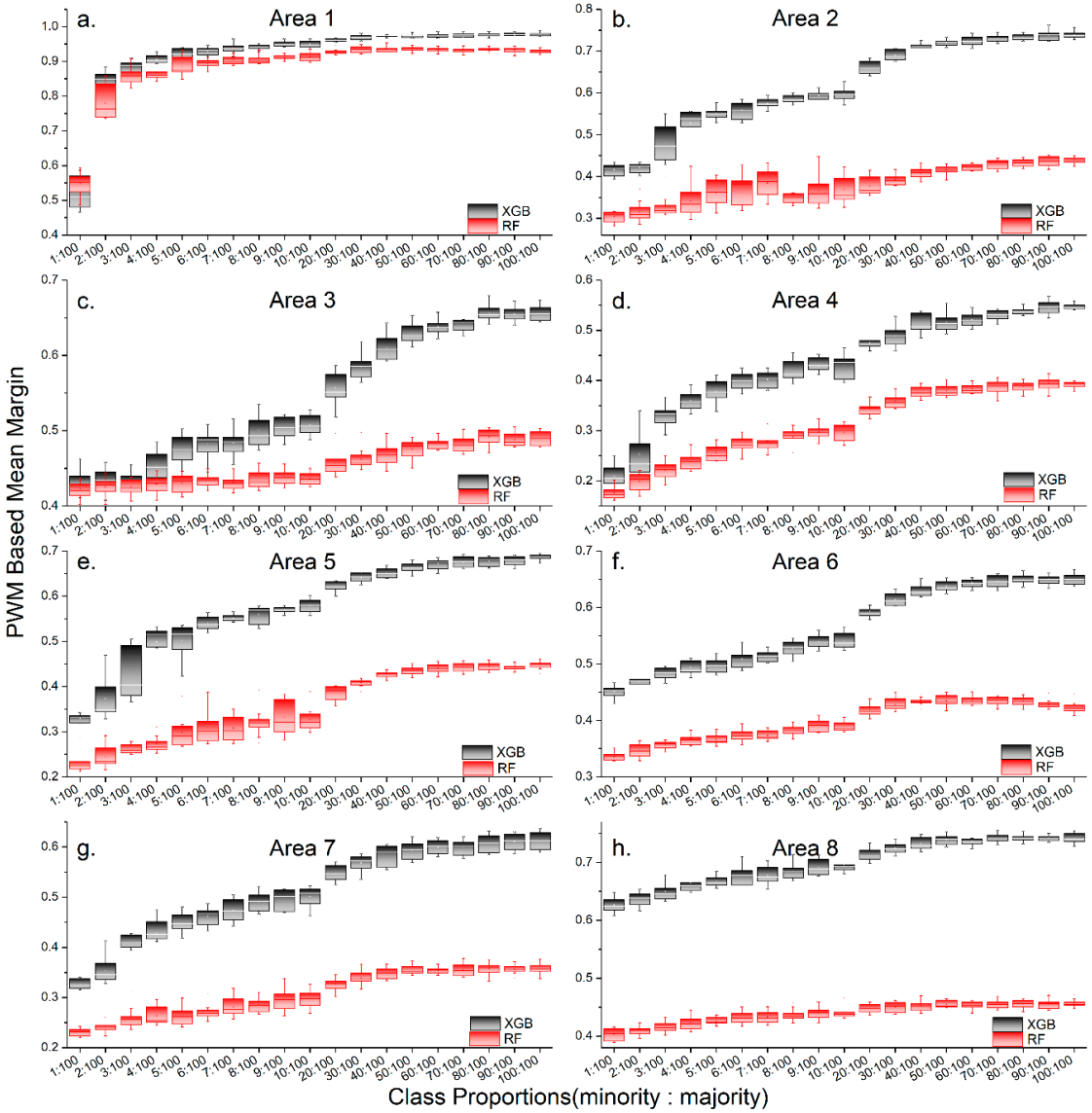

Figure 12. The boxplots for mean margin of XGB and RF using 10 independent trails across different minority proportions (minority: majority from 1:100 to 100: 100). The results of different areas are at: (**a**) area 1, (**b**) area 2, (**c**) area 3, (**d**) area 4, (**e**) area 5, (**f**) area 6, (**g**) area 7, (**h**) area 8.

Margin entropy, which measures the diversity of the PWM, is expected to be high but not close to its maximum. Only the margin entropies of XGB meet the expectations of this indicator (Figure 13). The margin entropy values by XGB for each area are more moderate, appropriate high but not close to the maximum value (3.32), comparing with values by RF. In detail, the margin entropies in area 1 by XGB are higher than that of RF; in region 2 to 7, the values by XGB range from 2.00 to 3.00 whereas the values of RF are close to the maximum value of 3.32. A high entropy value (close to 3.32) implies balanced diversity, which is not expected in ensemble prediction. Though the complex relationship between diversity and ensemble accuracy is still not well understood, we can only qualitatively compare the diversity between models.

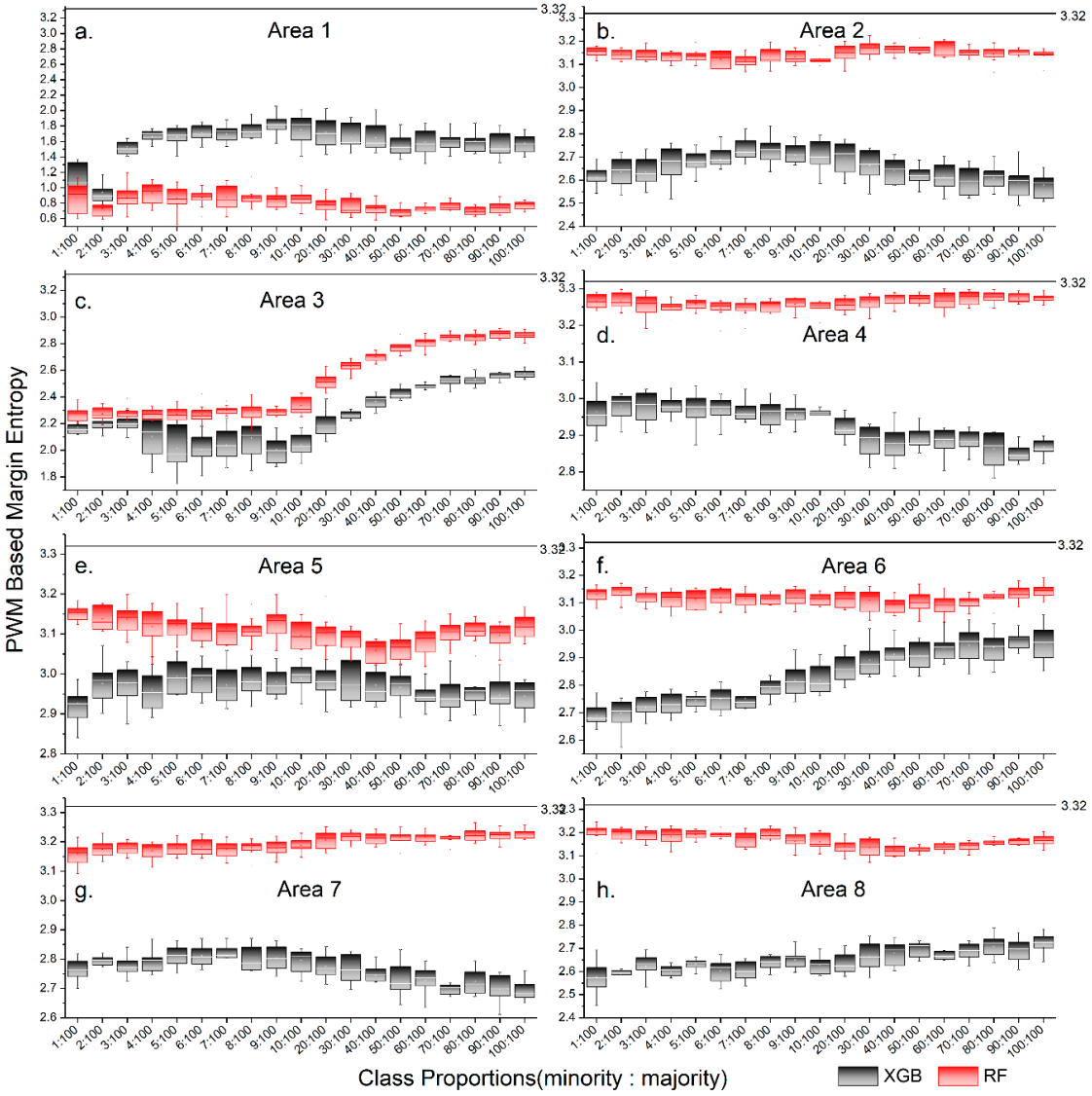

Figure 13. The boxplots of margin entropy of XGB and RF using 10 independent trails across different minority proportions (minority: majority from 1:100 to 100:100). The auxiliary line (y = 3.32) represents the theoretical maximum of the margin entropy with a 10-bins distribution. The results of different areas are at: (**a**) area 1, (**b**) area 2, (**c**) area 3, (**d**) area 4, (**e**) area 5, (**f**) area 6, (**g**) area 7, (**h**) area 8.

In general, XGB provides performance with reasonable accuracy and low uncertainty of correct predictions in VHR imbalanced LCC scenarios. Moreover, to obtain relatively stable classification accuracy, the samples of the minority class do not need to be increased to closely equal those of the majority class when using XGB. Appropriate sample imbalance can result in a cost-effective classification result in terms of accuracy and certainty.

3.2. Experiment 2: Influence of Minority Class Spectral Separability on XGB's Performance

Spectral separability is one of factors affecting the sample requirement. For Experiment 2, the tree and water classes in area 8 were assigned as the minority class respectively, presenting a striking contrast in spectral separability that affects remote sensing imbalanced learning. The results will be described according to the same aspects as in Section 3.1.

1. Overall performance. Since there are 11 classes in area 8, the contribution of each class for OA is more relatively limited in comparison with simple areas. Thus, the highest OA (appears when sample set is balanced and almost balanced) in both cases is almost unaffected, very close to 92%. However,

in an extremely imbalanced scenario, the largest difference in terms of OA between the two cases is approximately 5% (Figure 14). This finding indicates that the negative impact of sample imbalance on OA is also influenced by spectral separability, which has not been explicitly discussed in many studies. The influence of spectral separability can also be diminished by lower sample imbalance (almost equal OA after sample balance reaches minority:majority > 50:100). This is true for both XGB and RF, though the XGB performs better than RF. Finally, when the samples of each class are relatively balanced, the OA is gradually influenced by the spectral separability between classes (when assigning different minority classes, they achieve nearly the same OA).

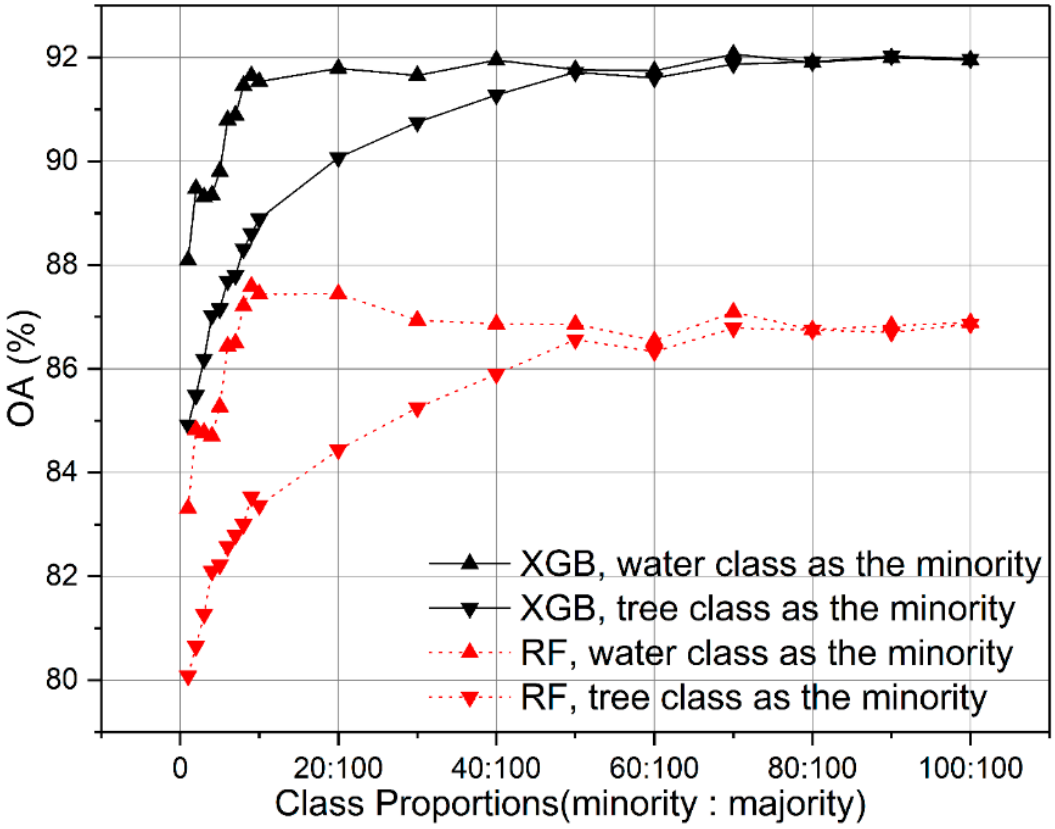

Figure 14. Overall accuracy of XGB and RF for the experiments when tree class and water class in area 8 are assigned to be the minority class in the training sets, respectively, across different minority proportions (minority: majority from 1:100 to 100:100).

Similarly to the result in Figure 5, the ED and SD performance is relatively stable in Figure 15a,b; that is to say, the sample imbalance mainly influences the overall QD. However, because the spectral separability of the water class from other classes is better than that of the tree class, this influence of imbalance on QD is not as variable as it is for the tree class. Thus, when investigating the effect of sample imbalance, common and good spectral separability lead to different error behaviours. However, when the water is the minority class, the three error components for both XGB and RF seem to be relatively stable; when tree is the minority, QD is obviously diminished and ED increases for both two methods. In addition, XGB has lower error components across different class imbalance than RF.

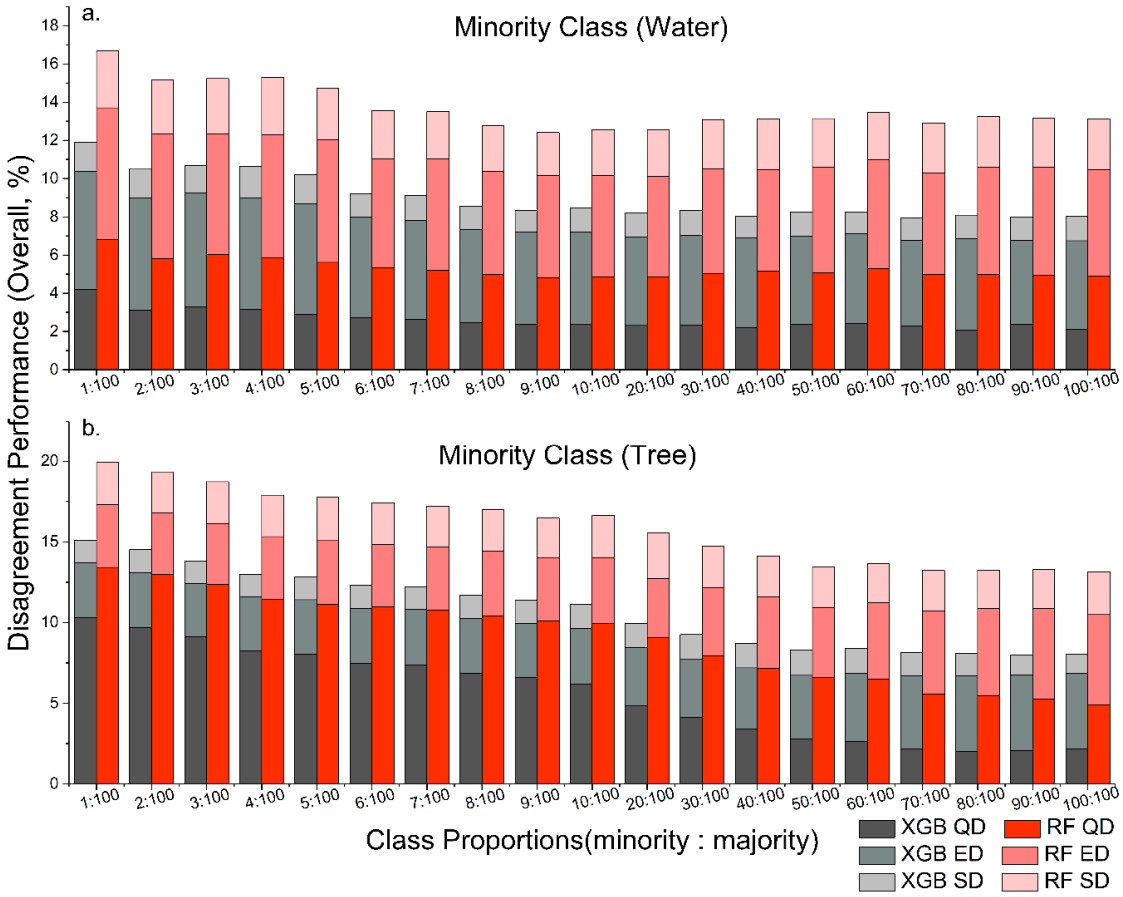

Figure 15. Overall disagreement performance of XGB and RF for area 8 when the minority class is of different spectral separability (**a**): water; (**b**): tree.

2. Recognition efficiency of the minority class. Visualization performances of XGB for the minority classes (water and tree) in area 8 are shown in Figure 16. As shown in Figure 16b–d), there is almost no significant difference. In fact, the water class includes 267,493 pixels in reference map (Figure 16a); XGB can distinguish most of the minority samples even with extremely imbalanced samples, reaching 93.42% when the training set imbalance ratio of minority:majority = 10:100, 94.97% at 50:100, and 95.76% at 100:100. In contrast, when the minority in the training dataset is the tree class, which has common spectral separability, the performance on the minority varies considerably. Only 19.44% of the tree pixels can be correctly identified with sample imbalance at 10:100. However, even if the sample amount has increased fivefold, the ratio increases by only 13.49%. With balanced training data, the accuracy for this minority is 67.01%. Compared with sample imbalance, recognition efficiency on the minority class in terms of accuracy is much more sensitive to spectral separability.

This conclusion is more clearly illustrated by the *F*1 score curves in Figure 17, where the curve of water is considerably higher than the corresponding results of tree. Moreover, the *F*1 score curves of water can quickly reach an approximate maximum when the minority proportion reaches 5% of the majority class, while the proportion is 50 to 90% for the tree class. Additionally, the *F*1 performance of XGB is much higher than RF when the minority class suffering common spectral separability (Figure 17, tree as the minority class) with class imbalance at 10:100 to 40:100. Therefore, the dominant factors affecting the per class accuracy are not always the same for different classes.

Similar to the *F*1 score curves (Figure 17), for both tree and water class, the producer and user accuracies of the minority class will still be greatly affected (almost 0) when the sample imbalance ratio is lower than 10:100 (Figure 18a). However, for class with a relatively common spectral separability, as samples become more and more imbalanced, user accuracy of tree class (Figure 18b) is greatly

reduced (consistent with Figure 7a). Comparing Figure 18a,b, after the imbalance ratio is larger than 10:100, the accuracy curves of the water class are stable, which is different from the curves of the tree class. This indicated that the sample imbalance has different effects on the recognition accuracy of the minority classed with different spectral separability.

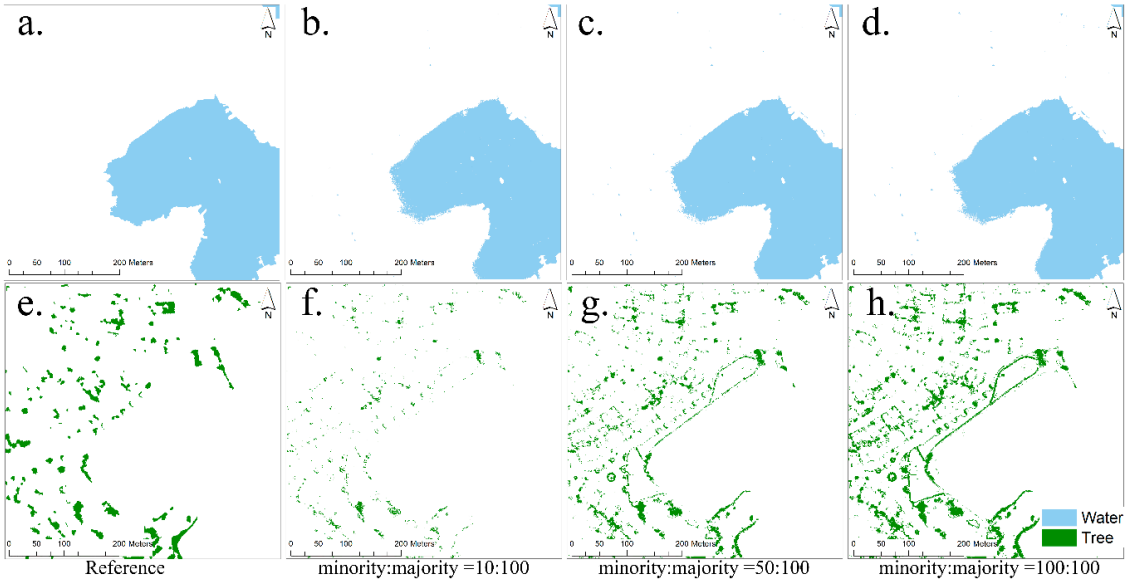

Figure 16. Minority class classification maps of XGB with different sample imbalance and spectral separability. The first row represents reference map (**a**) and the results when water class is the minority with different minority proportions (minority: majority) at: (**b**) 10:100, (**c**) 50:100, (**d**) 100:100. The second row represents the corresponding data (**e**: reference map for tree class) and results at: (**f**) 10:100, (**g**) 50:100, (**h**) 100:100 when tree class is the minority. Colours refer to different classes (water: blue, tree: green).

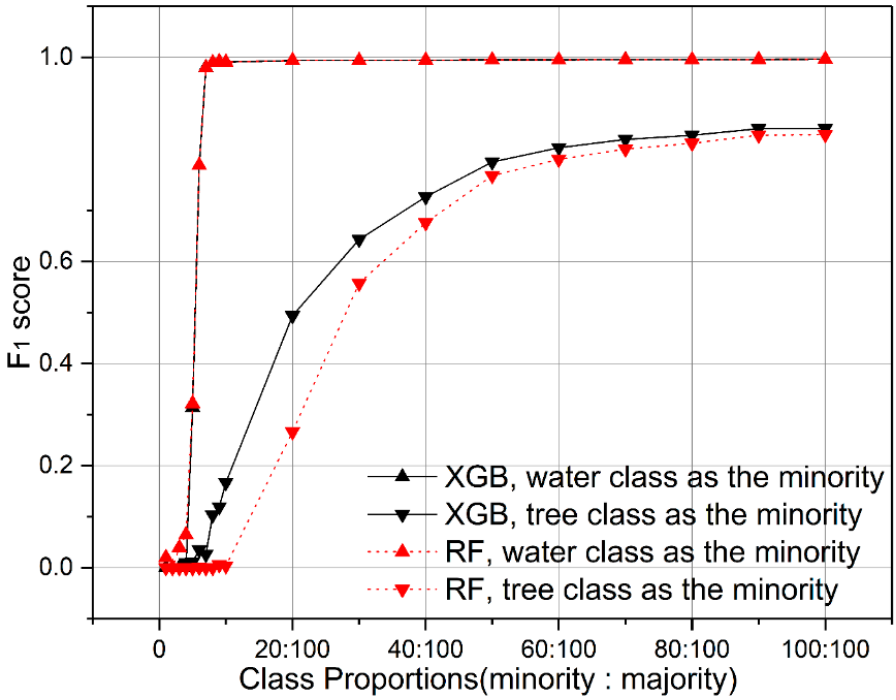

Figure 17. The F_1 score curves for the minority class of XGB and RF with different spectral separability of the minority class across different minority proportions (minority: majority from 1:100 to 100:100).

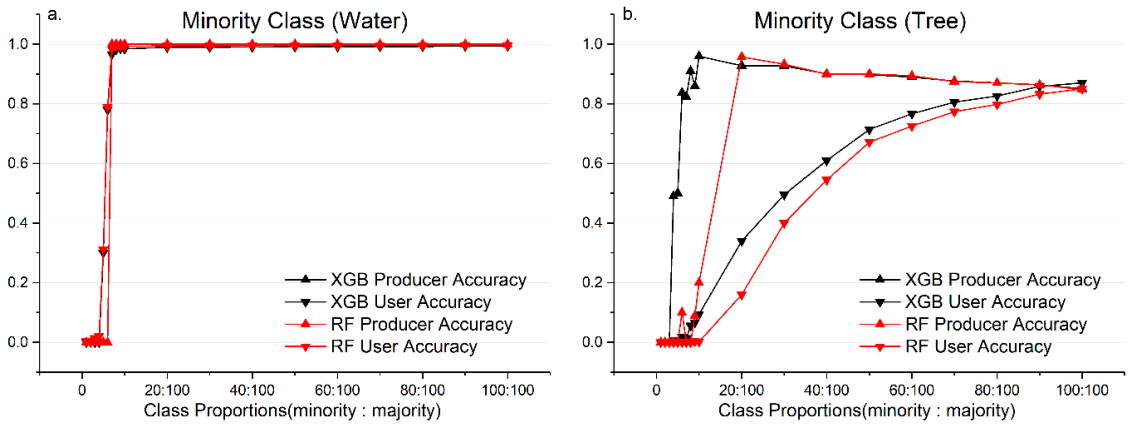

Figure 18. Curves of producer and user accuracies for area 8 when the minority class is of different spectral separability ((**a**): water; (**b**): tree) across different sample imbalance.

Obviously, the error components when water is the minority class are simply the QD, and the QD is relatively consistent after the sample imbalance reaches 20:100 (Figure 19). In the tree experiment, although the QD greatly improves with the decrease of sample imbalance, ED and SD appear and increase after the imbalance reaches 20:100, meaning that the tree instances are mixed with other classes in the prediction. This result also indicated that the behaviours of sample imbalance and spectral separability on the disagreement component are different. However, the performance of XGB and RF on the error component is quite close.

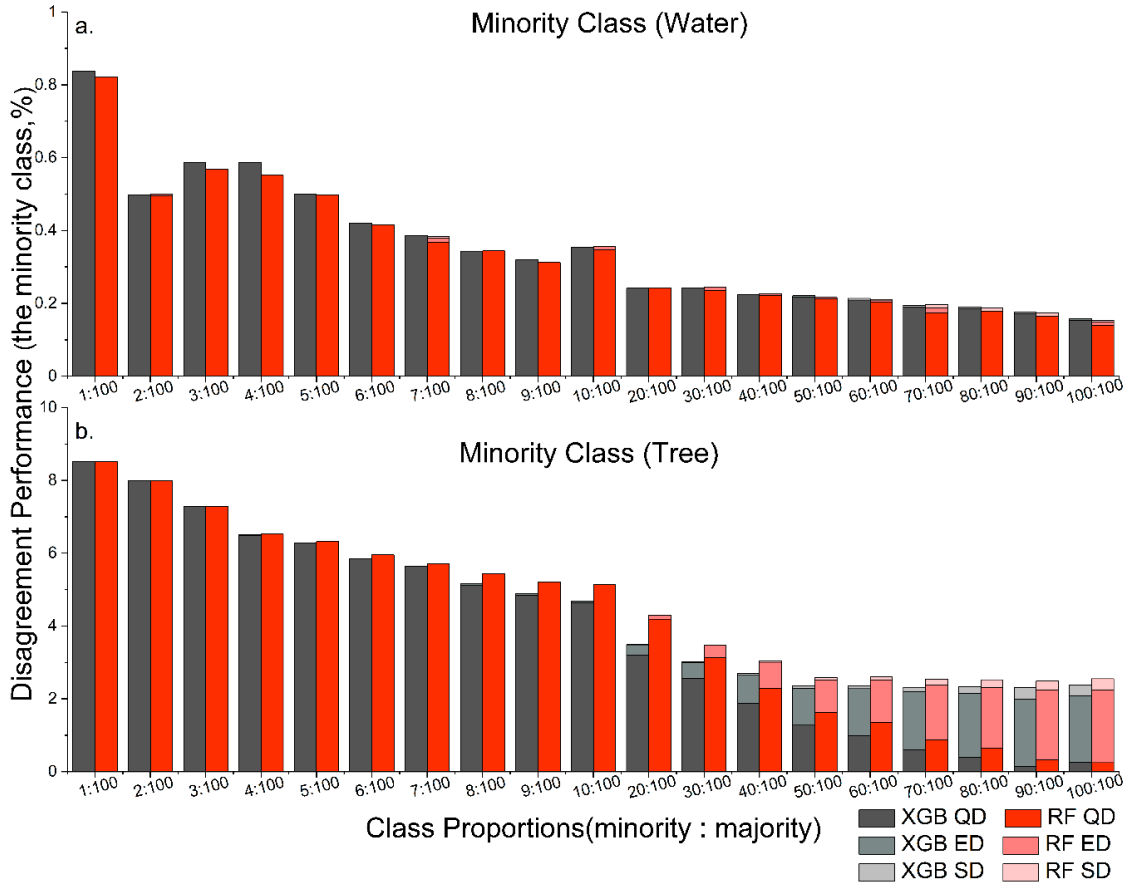

Figure 19. Disagreement performance of specific class using XGB and RF for area 8 when the minority class is of different spectral separability (**a**): water; (**b**): tree.

3. Performance uncertainty. In Figure 20, the cumulative frequency distribution curves for PWM with different minority class and sample imbalance are compared. When the sample imbalance is at minority:majority = 1:100, the PWM behaviour of XGB with water class as the minority class is obviously different from the result of tree class (Figure 20a). Meanwhile, as the relief of sample imbalance, leaf-shaped pairwise margin distributions are almost identical. That is, when the sample imbalance is satisfied a certain extent (such as minority:majority > 10:100 in our experiment), XGB is insensitive to spectral separability with regard to performance uncertainty measured by PWM. For the water class, although the correct classification exhibits a much higher margin even with extremely imbalanced samples, the incorrect predictions also tend to improve in PWM with the increased proportion of the minority class samples.

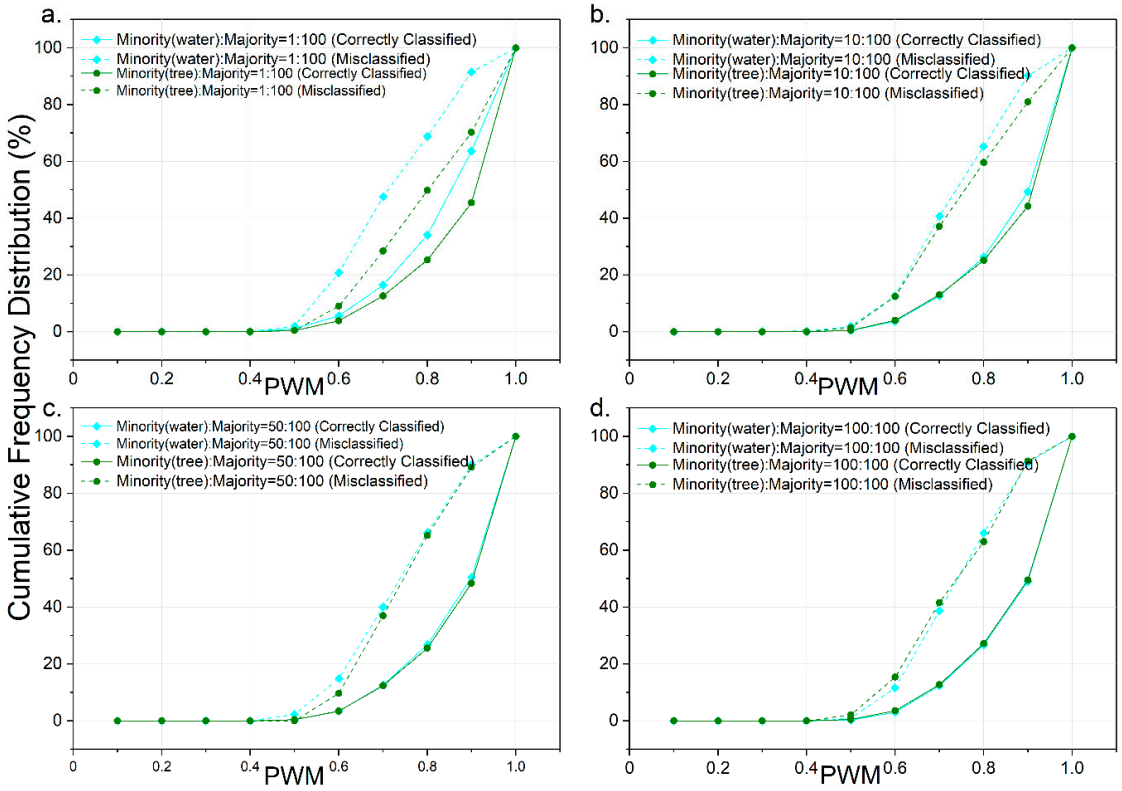

Figure 20. The PWM cumulative frequency distribution curves grouped with correctly and misclassified instance of XGB when tree (green) and water (blue) in area 8 are assigned to be the minority class in the training sets, respectively. The results of different sample imbalance are (minority: majority) at: (**a**) 1:100, (**b**) 10:100, (**c**) 50:100, (**d**) 100:100.

4. Discussion

Previous studies have demonstrated the effectiveness of RF under imbalanced learning conditions [1,58]; however, these studies provided only certain aspects of efficiency analysis of finite complex classification systems. [1] mainly used accuracies and the Kappa coefficient to evaluate the performance of RF on forest classification. [58] used the geometric mean and the F-score to evaluate the imbalance issues in a big data scenario with no attention to uncertainty. However, both of these previous studies focused on agreement evaluation and did not provide insight on the error components (disagreement performance). This study provides a comparison between XGB and RF in terms of both classification accuracy and performance uncertainty, not only focusing on agreement but also on disagreement (errors). Although both XGB and RF are decision tree-based ensemble algorithms, the accuracy assessment and error components are quite closed (XGB performs slightly better in

relatively complex areas), the margin performance associated with the correct classification of XGB is considerably higher than that of RF.

To explain this difference, we need to consider the different principles of XGB and RF. Given the same amount of the majority class samples, increasing the number of samples for the minority class provides continuous information for XGB as it iterates throughout the gradient boosting. However, for RF, which performs bootstrap resampling, the information assigned to each baseline classifier for model improvement is very limited. Thus, when the class proportion (minority:majority) is above 10:100 in the training set, XGB yields a relatively faster accuracy (overall and specific class) improvement than does RF as the minority class proportion increases. It is indicated that XGB can achieve a considerable performance with RF by using fewer samples.

Disagreement performance also brings new insights into the errors. Rarity samples make partial description of the class, inevitably resulting in a narrow statistical characterization of the reflectance [43]. This is the reason why extreme sample imbalance mainly influences the quantity disagreement for both the minority class and overall disagreement. However, as the imbalance fades, in a limited feature space, the characterization of the reflectance for each class may be overlapped; that is to say, similarity between classes increases. Therefore, the exchange and shift disagreement are much more obvious [43] for both XGB and RF when the samples of each class tend to be balanced. Thus, when using XGB for imbalanced learning, the appropriate sample imbalance can achieve a compromise between the classification benefit and the sample cost, both in terms of agreement and disagreement performance. All of the above discussion is applicable to a class with common spectral separability; undoubtedly, there is a quite complex relationship involving spectral separability, sample imbalance and model performance. Each factor has different importance to classification performance with different sample imbalances.

In terms of performance uncertainty, based on an analysis of the margin-weighted confuse matrices, XGB shows a reasonable level of improvement over RF regarding per-class uncertainty (the average PWM of correctly classified instances is 46% larger and 90% associated with incorrect classifications, though the former is what we expect, the latter is not). This result is also consistent with previous research [36], which showed that boosting was the most effective method for improving margin. The differences in the margin evaluation between bagging (RF) and boosting (XGB) models can be simply explained using bias-variance decomposition [36]. For bagging methods such as RF, bootstrap resampling is expected to provide an ideal bagging set, which consists of approximately truly independent samples. However, this expectation is difficult to satisfy, especially for imbalanced datasets. Thus, the RF baselines have certain correlations, i.e., they are not absolutely identical and independent; consequently, RF reduces the variance but not the bias (the errors that baselines make for baseline hypotheses) directly, resulting in its poor margin evaluation performance. In contrast, boosting methods, such as XGB, do more than simply reducing the variance. With each iterative correction of the gradient residual, the XGB method reduces both the bias and the variance [36]. However, the training samples must be sufficient to perform a reasonable number of gradient boosting iterations in XGB construction. When there are not enough samples, XGB cannot obtain adequate gradient corrections, resulting in poor uncertainty. This is why the PWM statistic is variously affected when the training data are extremely imbalanced. Due to the model construction mechanisms, iteration of XGB causes the model to continuously perform residual corrections on the wrong behaviour, while the bagging (RF) approach does not modify the errors. The residual correction operations provide the output with more confidence. However, because the gradient residual correction is also effective for misclassified instances, the PWM values of these instances provided by XGB are also relatively high (Table 3). It is a drawback needed to be improved when using XGB to solve imbalance learning for remote sensing data.

In addition, from Figure 2, we can clearly see that the tree and farmland with crop cover classes in Beihai overlap significantly in the DN distribution. Thus, the minority class (tree) results in numerous misclassifications as crop cover, which leads to a loss of accuracy in the overall performance for area 3. This is why the results for area 3 are not as significant as those for the rest of the areas (Figure 4).

5. Conclusions

This study presented an insight of imbalance learning using XGB with remote sensing VHR data in terms of the overall performance, the accuracy on the minority class, and the uncertainty. The comparative experiments are conducted on 8 study areas with different complexities. The results demonstrate that: (1) In VHR image classification, XGB performs better than RF under different imbalanced learning scenarios, and this advantage is especially obvious in complex study areas; (2) Imbalanced samples (minority:majority > 10:100) mainly influence the user accuracy of the minority class with common spectral separability. However, the negative impact on XGB is less than that on RF; (3) The average PWM of correct classification provided by XGB is 0.82, which is about 46.30% higher than the RF, indicating that XGB has lower performance uncertainty. Moreover, the performance uncertainty of XGB is insensitive to spectral separability when sample imbalance is satisfied a certain extent (minority:majority > 10:100 in our experiment). Therefore, XGB is an effective method for imbalanced land cover classification from VHR images. Additionally, appropriate sample imbalance helps to improve the trade-off between accuracy of XGB and the sample cost. In the future, we will further improve the performance of XGB on extremely imbalanced remote sensing data classification, and multiple minority classes in remote sensing land cover classification.

Author Contributions: Fei Sun and Run Wang designed the experiment; Fei Sun and Bo Wan collected the required data; Fei Sun and Youxin Huang conducted the analysis of the results; and Fei Sun, Run Wang, Bo Wan, Yanjun Su, Qinghua Guo and Xincai Wu contributed towards writing the manuscript.

Funding: This work is partially funded by the National Key Research & Development (R&D) Plan of China under Grant 2017YFB0503600 and the National Natural Science Foundation of China under Grant 41674100.

Acknowledgments: The authors would like to thank the Land and Resources Bureau of Beihai city for providing the aerial datasets; Lingfeng Yuan for help with manual validation; and the anonymous reviewers for their constructive comments and suggestions, which helped improve our paper.

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

Appendix A

1. Brief introduction of margin-based certainty margin based confidence measures

Mean margin (μ) can be defined as follows:

$$\mu = \frac{n_c\mu_c - n_m\mu_m}{n_c + n_m} \tag{A1}$$

where nc and nm are the amount of correct and wrong classification, where μc and μm are the average margins of corresponding PWM.

Margin entropy (H) are calculated by using margin bins histogram. First, divide the margin distribution of the test set into n bins (10 in this paper). Then calculate the probabilities (Pr) of each bin. Then use the Shannon's formula to calculate this margin entropy, as follows:

$$H = -\sum_{i=0}^{n-1}\left[\Pr(m_i) \times \log_2 \Pr(m_i)\right], \Pr(m_i) \neq 0 \tag{A2}$$

2. Brief introduction of disagreement performance.

In this paper, the disagreement performance, 3 error components, quantity disagreement (QD), exchange disagreement (ED) and shift disagreement (SD), are calculated using confusion matrix. C_{ij} represents the element of the i-th row and j-th column in the confusion matrix. So, the error components of class j can be defined as:

$$d_j = \left(\sum_{i=1}^{J}\left(C_{ij} + C_{ji}\right) - 2 \times C_{jj}\right) \times 100\% / \sum_{i=1}^{J}\sum_{j=1}^{J} C_{ij} \tag{A3}$$

$$QD_j = \left|\sum_{i=1}^{J}\left(C_{ij} - C_{ji}\right)\right| \times 100\% / \sum_{i=1}^{J}\sum_{j=1}^{J} C_{ij} \tag{A4}$$

$$ED_j = 2 \times \left[\sum_{i=1}^{J} MINIMUM\left(C_{ij}, C_{ji}\right) - C_{jj} \right] \times 100\% / \sum_{i=1}^{J} \sum_{j=1}^{J} C_{ij} \quad \text{(A5)}$$

$$SD_j = d_j - q_j - e_j \quad \text{(A6)}$$

where d_j is the difference of class j. For mutil-class scenario, the overall disagreement performance can be defined as:

$$QD = \sum_{j=1}^{J} QD_j/2 \quad \text{(A7)}$$

$$ED = \sum_{j=1}^{J} ED_j/2 \quad \text{(A8)}$$

$$SD = \sum_{j=1}^{J} SD_j/2 \quad \text{(A9)}$$

More details can be found in [1,42,43].

3. Reference map and classification maps of area 8 with different imbalanced samples.

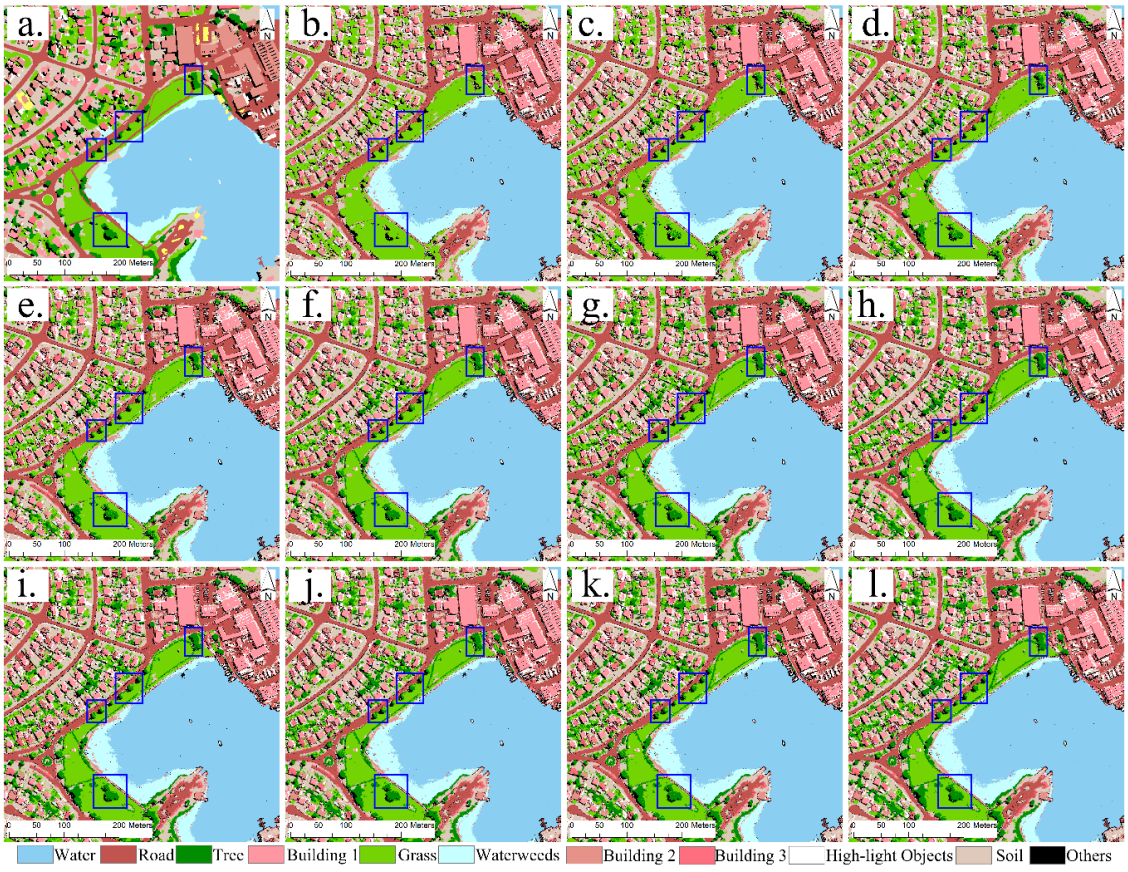

Figure A1. Reference map (**a**) and Classification maps of XGB using the training sets with the ratio of minority: majority at: (**b**) 1:100, (**c**) 10:100, (**d**) 20:100, (**e**) 30:100, (**f**) 40:100, (**g**) 50:100, (**h**) 60:100, (**i**) 70:100, (**j**) 80:100, (**k**) 90:100, (**l**) 100:100, for area 8.

References

1. Mellor, A.; Boukir, S.; Haywood, A.; Jones, S. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 155–168. [CrossRef]
2. Mellor, A.; Boukir, S. Exploring diversity in ensemble classification: Applications in large area land cover mapping. *ISPRS J. Photogramm. Remote Sens.* **2017**, *129*, 151–161. [CrossRef]

3. Foody, G.M. Status of land cover classification accuracy assessment. *Remote Sens. Environ.* **2002**, *80*, 185–201. [CrossRef]
4. Geiß, C.; Pelizari, P.A.; Marconcini, M.; Sengara, W.; Edwards, M.; Lakes, T.; Taubenböck, H. Estimation of seismic building structural types using multi-sensor remote sensing and machine learning techniques. *ISPRS J. Photogramm. Remote Sens.* **2015**, *104*, 175–188. [CrossRef]
5. Lippitt, C.D.; Rogan, J.; Li, Z.; Eastman, J.R.; Jones, T.G. Mapping selective logging in mixed deciduous forest: A comparison of Machine Learning Algorithms. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 1201–1211. [CrossRef]
6. Leichtle, T.; Geiß, C.; Lakes, T.; Taubenböck, H. Class imbalance in unsupervised change detection—A diagnostic analysis from urban remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *60*, 83–98. [CrossRef]
7. Foody, G.M.; Mathur, A.; Sanchez-Hernandez, C.; Boyd, D.S. Training set size requirements for the classification of a specific class. *Remote Sens. Environ.* **2006**, *104*, 1–14. [CrossRef]
8. Foster, P. Machine Learning from Imbalanced Data Sets 101 (Extended Abstract). In Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets, Austin, TX, USA, 31 July 2000.
9. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
10. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
11. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
12. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
13. Krawczyk, B.; Woźniak, M.; Schaefer, G. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Appl. Soft Comput.* **2014**, *14*, 554–562. [CrossRef]
14. Ha, J.; Lee, J.-S. A New Under-Sampling Method Using Genetic Algorithm for Imbalanced Data Classification. In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, Danang, Vietnam, 4–6 January 2016; pp. 1–6.
15. Nekooeimehr, I.; Lai-Yuen, S.K. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Syst. Appl.* **2016**, *46*, 405–416. [CrossRef]
16. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* **2002**, *16*, 321–357. [CrossRef]
17. Andrew, E.; Taeho, J.; Nathalie, J. A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Comput. Intell.* **2004**, *20*, 18–36. [CrossRef]
18. Wang, B.; Pineau, J. Online Bagging and Boosting for Imbalanced Data Streams. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3353–3366. [CrossRef]
19. Krawczyk, B.; Galar, M.; Jeleń, Ł.; Herrera, F. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Appl. Soft Comput.* **2016**, *38*, 714–726. [CrossRef]
20. Hassan, A.K.I.; Abraham, A. Modeling Insurance Fraud Detection Using Imbalanced Data Classification. In *Advances in Nature and Biologically Inspired Computing*; Springer: Berlin, Germany, 2016; pp. 117–127.
21. López, V.; del Río, S.; Benítez, J.M.; Herrera, F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets Syst.* **2015**, *258*, 5–38. [CrossRef]
22. Wu, D.; Wang, Z.; Chen, Y.; Zhao, H. Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. *Neurocomputing* **2016**, *190*, 35–49. [CrossRef]
23. Bruzzone, L.; Serpico, S.B. Classification of imbalanced remote-sensing data by neural networks. *Pattern Recognit. Lett.* **1997**, *18*, 1323–1328. [CrossRef]
24. Li, F.; Li, S.; Zhu, C.; Lan, X.; Chang, H. Cost-Effective Class-Imbalance Aware CNN for Vehicle Localization and Categorization in High Resolution Aerial Images. *Remote Sens.* **2017**, *9*, 494. [CrossRef]
25. Chen, X.; Fang, T.; Huo, H.; Li, D. Semisupervised Feature Selection for Unbalanced Sample Sets of VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 781–785. [CrossRef]
26. Graves, J.S.; Asner, P.G.; Martin, E.R.; Anderson, B.C.; Colgan, S.M.; Kalantari, L.; Bohlman, A.S. Tree Species Abundance Predictions in a Tropical Agricultural Landscape with a Supervised Classification Model and Imbalanced Data. *Remote Sens.* **2016**, *8*, 161. [CrossRef]

27. Loyola-González, O.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; García-Borroto, M. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing* **2016**, *175*, 935–947. [CrossRef]
28. Tien Bui, D.; Pradhan, B.; Lofman, O.; Revhaug, I. Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models. *Math. Probl. Eng.* **2012**, *2012*, 1–26. [CrossRef]
29. Pal, M.; Mather, P.M. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens. Environ.* **2003**, *86*, 554–565. [CrossRef]
30. 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.
31. Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F.; Yu, X.; Lu, X.; Xiang, Y. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Convers. Manag.* **2018**, *164*, 102–111. [CrossRef]
32. Carmona, P.; Climent, F.; Momparler, A. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *Int. Rev. Econ. Finance* **2019**, *61*, 304–323. [CrossRef]
33. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141. [CrossRef]
34. Rokach, L.; Schclar, A.; Itach, E. Ensemble methods for multi-label classification. *Expert Syst. Appl.* **2014**, *41*, 7507–7523. [CrossRef]
35. Bi, J.; Zhang, C. An Empirical Comparison on State-of-the-art Multi-class Imbalance Learning Algorithms and A New Diversified Ensemble Learning Scheme. *Knowl. Based Syst.* **2018**, *158*, 81–93. [CrossRef]
36. Schapire, R.E.; Freund, Y.; Barlett, P.; Lee, W.S. Boosting the margin: A new explanation for the effectiveness of voting methods. In Proceedings of the 14th International Conference on Machine Learning (ICML '97), Nashville, TN, USA, 8–12 July 1997; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA; pp. 322–330.
37. Panuju, D.R.; Paull, D.J.; Trisasongko, B.H. Combining Binary and Post-Classification Change Analysis of Augmented ALOS Backscatter for Identifying Subtle Land Cover Changes. *Remote Sens.* **2019**, *11*, 100. [CrossRef]
38. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Wolff, E. Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 607–611. [CrossRef]
39. Ustuner, M.; Balik Sanli, F. Polarimetric Target Decompositions and Light Gradient Boosting Machine for Crop Classification: A Comparative Evaluation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 97. [CrossRef]
40. Branco, P.; Torgo, L.; Ribeiro, R. A Survey of Predictive Mo delling under Imbalanced Distributions. CoRR. *arXiv* **2015**, arXiv:1505.01658.
41. Chawla, N.V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2010; pp. 875–886.
42. Pontius, R.G.; Millones, M. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* **2011**, *32*, 4407–4429. [CrossRef]
43. Pontius, R.G.; Santacruz, A. Quantity, exchange, and shift components of difference in a square contingency table. *Int. J. Remote Sens.* **2014**, *35*, 7543–7554. [CrossRef]
44. Guo, Q.; Li, W.; Liu, D.; Chen, J. A Framework for Supervised Image Classification with Incomplete Training Samples. *Photogramm. Eng. Remote Sens.* **2012**, *78*, 595–604. [CrossRef]
45. Madonsela, S.; Cho, M.A.; Ramoelo, A.; Mutanga, O.; Naidoo, L. Estimating tree species diversity in the savannah using NDVI and woody canopy cover. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *66*, 106–115. [CrossRef]
46. McGarigal, K.; Cushman, S.A.; Ene, E. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Available online: http://www.umass.edu/landeco/research/fragstats/fragstats.html (accessed on 1 May 2019).
47. Song, C.; Woodcock, C.E.; Seto, K.C.; Lenney, M.P.; Macomber, S.A. Classification and Change Detection Using Landsat TM Data: When and How to Correct Atmospheric Effects? *Remote Sens. Environ.* **2001**, *75*, 230–244. [CrossRef]

48. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *3*, 610–621. [CrossRef]
49. Li, W.; Guo, Q.; Elkan, C. A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 717–725. [CrossRef]
50. Richards, J.A. *Remote Sensing Digital Image Analysis*; Springer: Berlin, Germany, 1999.
51. García Nieto, P.J.; García–Gonzalo, E.; Arbat, G.; Duran–Ros, M.; Ramírez de Cartagena, F.; Puig-Bargués, J. Pressure drop modelling in sand filters in micro-irrigation using gradient boosted regression trees. *Biosyst. Eng.* **2018**, *171*, 41–51. [CrossRef]
52. Chen, L.; Zhang, T.; Li, T. Gradient boosting model for unbalanced quantitative mass spectra quality assessment. In Proceedings of the 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Shenzhen, China, 15–17 December 2017; IEEE: Piscataway, NJ, USA; pp. 394–399.
53. He, H.; Zhang, W.; Zhang, S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Syst. Appl.* **2018**, *98*, 105–117. [CrossRef]
54. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
55. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [CrossRef]
56. Cheng, F.; Zhang, J.; Wen, C.; Liu, Z.; Li, Z. Large cost-sensitive margin distribution machine for imbalanced data classification. *Neurocomputing* **2017**, *224*, 45–57. [CrossRef]
57. Kuncheva, L.I.; Whitaker, C.J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* **2003**, *51*, 181–207. [CrossRef]
58. Del Río, S.; López, V.; Benítez, J.M.; Herrera, F. On the use of MapReduce for imbalanced big data using Random Forest. *Inf. Sci.* **2014**, *285*, 112–137. [CrossRef]