



Article

Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique

Mengying Lin ¹ , Xuefen Zhu ^{1,*}, Teng Hua ¹, Xinhua Tang ¹, Gangyi Tu ² and Xiyuan Chen ¹ 

¹ Key Laboratory of Micro-Inertial Instrument and Advanced Navigation Technology of Ministry of Education, Sipailou Campus, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China; linmengying@seu.edu.cn (M.L.); 220203499@seu.edu.cn (T.H.); xinhua.tang@seu.edu.cn (X.T.); chxiyuan@seu.edu.cn (X.C.)

² School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Ning Liu Road, Nanjing 210044, China; tugangyi@nuist.edu.cn

* Correspondence: zhuxuefen@seu.edu.cn; Tel.: +86-136-4516-1372

Abstract: Ionospheric scintillation frequently occurs in equatorial, auroral and polar regions, posing a threat to the performance of the global navigation satellite system (GNSS). Thus, the detection of ionospheric scintillation is of great significance in regard to improving GNSS performance, especially when severe ionospheric scintillation occurs. Normal algorithms exhibit insensitivity in strong scintillation detection in that the natural phenomenon of strong scintillation appears only occasionally, and such samples account for a small proportion of the data in datasets relative to those for weak/moderate scintillation events. Aiming at improving the detection accuracy, we proposed a strategy combining an improved eXtreme Gradient Boosting (XGBoost) algorithm by using the synthetic minority, oversampling technique and edited nearest neighbor (SMOTE-ENN) resampling technique for detecting events imbalanced with respect to weak, medium and strong ionospheric scintillation. It outperformed the decision tree and random forest by 12% when using imbalanced training and validation data, for tree depths ranging from 1 to 30. For different degrees of imbalance in the training datasets, the testing accuracy of the improved XGBoost was about 4% to 5% higher than that of the decision tree and random forest. Meanwhile, the testing results for the improved method showed significant increases in evaluation indicators, while the recall value for strong scintillation events was relatively stable, above 90%, and the corresponding F1 scores were over 92%. When testing on datasets with different degrees of imbalance, there was a distinct increase of about 10% to 20% in the recall value and 6% to 11% in the F1 score for strong scintillation events, with the testing accuracy ranging from 90.42% to 96.04%.

Keywords: GNSS; ionospheric scintillation detection; XGBoost; SMOTE-ENN



Citation: Lin, M.; Zhu, X.; Hua, T.; Tang, X.; Tu, G.; Chen, X. Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique. *Remote Sens.* **2021**, *13*, 2577. <https://doi.org/10.3390/rs13132577>

Academic Editor: Michael E. Gorbunov

Received: 1 June 2021

Accepted: 29 June 2021

Published: 1 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The ionosphere, the atmosphere at about 60 to 1000 km from the ground, is modulated by the ionizing effects of solar radiation, particle precipitation and the geomagnetic field. There are some typical ionospheric phenomena, such as the equatorial ionization anomaly (EIA) and equatorial plasma bubbles at low altitudes [1–3], as well as the tongue of ionization at high latitudes [4,5], by which radio waves such as global navigation satellite system (GNSS) signals may be severely affected. When passing through the ionospheric irregularities, the signals are plagued with rapid fluctuation, phase shifts, delay, multipath, and even loss of tracking loop. More seriously, this phenomenon of ionospheric scintillation occurs more frequently and severely in both low-latitude regions and polar regions, compromising positioning accuracy and continuity [6,7]. In high-latitude regions, the occurrence of ionospheric scintillation is more apparent during geomagnetic storms, and the formation of irregular structures and ionospheric scintillation activities appears to

be related to the magnetic local time, resulting in the fast movement of large-scale ionospheric plasma structures and phase fluctuations [8–10]. In low-latitude areas, equatorial scintillation activities are stronger and frequent during years with solar maxima, which are associated with the solar activities [11]. The GNSS receiver is affected by amplitude and phase fluctuation, carrier Doppler jitter, cycle slips, and the serious loss of lock and even navigation interruption, which have adverse effects on both ionospheric research and navigation and positioning services [12,13].

Evidence suggests that ionospheric scintillation plays a vital role in receiver performance in positioning and navigation services. To improve the quality and reliability of GNSS observations, the implementation of the automatic detection and monitoring of ionospheric scintillation is vital, especially during strong scintillation. Relevant ionospheric scintillation monitoring receivers (ISMRs) have recently been designed for use in commerce and research, which are used by presetting thresholds for amplitude and phase indices [14]. However, detection methods such as traditional thresholding technologies are influenced by various factors. To be specific, this approach requires detrending and filtering algorithms for observations and overlooks higher-order moment information of the GNSS signals [15]. Moreover, it is sensitive to false alarms due to factors such as multipath, and has been proved to present a lower detection accuracy of 81%, compared to the approach of manual visual inspection [15,16]. Some wavelet decomposition and transform-based techniques using Butterworth filters with non-indices are proposed as alternatives to overcome the problem of detrending, but they rely on expensive computation and complex implementation [17,18].

In recent years, researchers have investigated a variety of machine learning approaches to achieve automatic scintillation detection. Jiao et al. (2017) [19,20] firstly exploited the support vector machine (SVM) for amplitude scintillation detection on two classes. They used a mass of real data and manual labels in the training process and achieved a detection accuracy of 91–96%, outperforming other traditional triggering approaches and non-index techniques. A similar detection method for phase scintillation presents an accuracy of around 92% [21,22]. Lin et al. (2020) [23] analyzed the effects of a binary classification SVM model on hyperparameters and achieved excellent performance in testing. However, based on features such as scintillation indices and relevant maximum and average values, these methods require a filtering process for the scintillation indices S4/SigmaPhi and predetermined elevation mask of 30°. Ludwig-Barbosa, V. et al. (2021) [24] trained SVM models with features combining amplitude/phase scintillation indices along with corresponding maximum and mean values, as well as intensity power spectral density (PSD), showing about 91% accuracy in the detection of ionospheric scintillation. Besides the detection of ionospheric scintillation by an SVM, similar research such as that on jamming detection in GNSS bands with an SVM was conducted, with 94.4% accuracy [25]. An accuracy of 91.36% was found when performing a similar task using convolutional neural networks [26], which are widely researched for various classification tasks, showing great performance [26–28]. Linty et al. (2019) [16] proposed a decision tree algorithm relying on the in-phase and quadrature correlator outputs of the receiver tracking loop, which are considered as sample features after simple computation. The results of 10-fold cross-validation show that the accuracy of amplitude scintillation detection reached 96.7% for features consisting of S4, the carrier-to-noise rate (C/N0) and satellite elevation. Furthermore, for features used by correlator outputs and corresponding combinations, the cross-validation accuracy increased from 98.0% to 99.7% when using a random forest algorithm. The overall F1 score reached a high value of 90%, compared with the value of 80.1% achieved with the semi-hard rule [16]. Based on this research, Franzese et al. (2020) [29] proposed semi-supervised scintillation detection using the DeepInformax technique, which presents a validation accuracy in accord with that of the decision tree model. The decision tree and random forest models are vulnerable to the problem of overfitting if the high model complexity is designed improperly. More importantly, these methods mentioned above rarely consider the imbalanced phenomenon of different intensities of scintillation events [16,19–25,29].

Insufficient data related to strong scintillation may cause great trouble in model training. Moreover, the events to be tested with different intensities may show different degrees of imbalance. Such imbalance in different intensities of scintillation events is rarely taken into consideration. Among the occasional ionospheric scintillation events, strong scintillation is a minority class compared to weak and moderate scintillation. For this class of data, few classification algorithms can accurately describe the inherent characteristics due to the lack of information about the minority class, which causes the decision boundary to be greatly compressed in the classification model [30]. Although the overall accuracy of the final model is relatively high, there is some missed detection for strong scintillation events. It can be optimized by solving the imbalance of the dataset to reduce the problem of missing detection points for strong scintillation events in the testing set.

In this paper, an improved machine learning method is proposed for improving the automatic detection accuracy for strong ionospheric scintillation, and it is compared in detail with the present decision tree and random forest, which show high accuracy for the validation set [16]. There is also a brief comparison with an SVM and CNN, which have recently been utilized in relevant research [19–28]. The proposed approach can provide significant guidance for designing robust GNSS receivers as well as research on atmospheric layers and space weather. Several aspects of this approach can be described as follows:

- (1) To propose a high-performance detection method for XGBoost based on the decision tree algorithm, assuring good overall detection accuracy for three intensities of ionospheric scintillation;
- (2) To compare with decision tree and random forest based on cross-validation, proving the superior accuracy of the XGBoost algorithm and mitigation of the overfitting problem;
- (3) To compare different resampling techniques for an imbalanced dataset consisting of a majority of weak/medium scintillation events and a minority of strong scintillation events, proving the great performance of SMOTE-ENN according to evaluation indicators;
- (4) To make a brief overall comparison with the decision tree, random forest, SVM and CNN, evaluating the performance in terms of accuracy, computational load and applicability, focusing on detailed comparison with the first two methods with high performance hereafter;
- (5) To evaluate the performance of the proposed improved method on different degrees of imbalanced training datasets and testing datasets, respectively, proving the effectiveness in enhancing the detection accuracy for strong scintillation events.

In this work, a strategy of using data processing technology for imbalanced ionospheric scintillation events is proposed, and we applied our strategy to space weather detection for the first time, which successfully improved the detection accuracy for strong scintillation events. Additionally, an improved machine learning algorithm with proven high overall accuracy for scintillation events of different intensities is proposed. The general overviews are described in Section 2, including an introduction to ionospheric scintillation, the data collection system and feature extraction. Section 3 presents an improved machine learning method, giving a theoretical introduction to the XGBoost algorithm and SMOTE-ENN technique. Section 4 validates the optimal detection performance of the proposed approaches, showing and evaluating the quantitative results obtained from the training and testing on datasets with different degrees of imbalance. Section 5 discusses various training and testing cases, and Section 6 draws the main conclusions.

2. General Overviews

This section presents brief descriptions of GNSS ionospheric scintillation, the data collection system and the feature extraction process.

2.1. GNSS Ionospheric Scintillation

GNSS satellites transmit low-rate navigation messages modulated by unique spreading sequences or codes, which are carried by radio frequency (RF) signals and received by GNSS receivers on the Earth. Corresponding computations on position, velocity and time can be utilized in land applications such as autonomous vehicle navigation and vehicle tracking monitoring, in marine applications such as ocean transportation and inland river shipping, and in aviation applications such as route navigation, airport scene monitoring and precision approach.

In the propagation of RF signals from satellites to receivers, there are natural disturbances in the GNSS signals when they pass through the ionospheric irregularity, resulting in temporal delay and fluctuations, degrading the accuracy, integrity and reliability of the system's performance. Figure 1 shows the effects of ionospheric scintillation on the process of transiting satellite signals to GNSS receivers. Ionospheric scintillation, a form of major interference, is difficult to model due to its quasi-random nature. With the increased demand for and dependence on navigation systems in various fields, research on improving the accuracy of scintillation detection is significant.

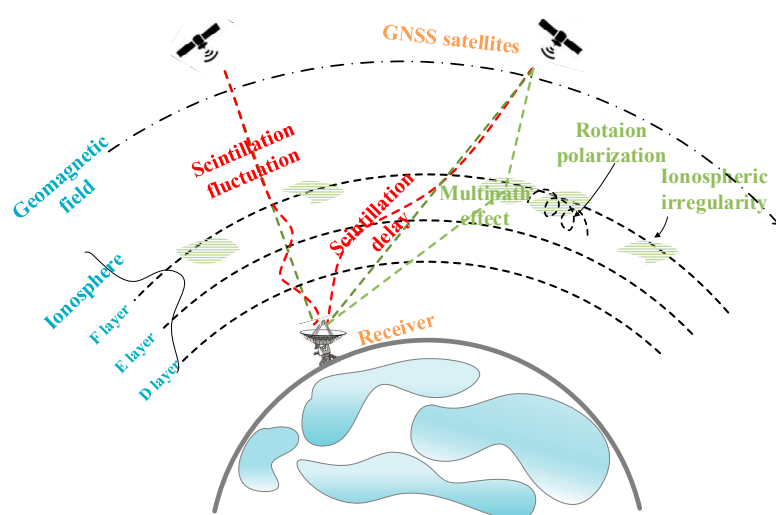


Figure 1. Representative effects of ionospheric scintillation on GNSS receiver. The green blocks refer to the ionospheric irregularities on the area of ionospheric layer, and the red lines represent the main disturbances on the ground receiver.

Ionospheric scintillation is classified into amplitude scintillation and phase scintillation; the former is more frequent in low-latitude areas. It is extremely important for a better detection method to be developed for the Beidou navigation system (BDS), which has been built in recent years with an immature scintillation detection technique [31]. Based on existing GPS data consisting of strong scintillation that were collected in the latest peak year for solar activity, relevant research on scintillation detection can be carried out to develop corresponding technology for the BDS. The performance can also be evaluated in the coming peak year for solar activity (2023–2024). S4, commonly used as an indicator, ranges from 0 to 1, for which a large value represents a stronger intensity of ionospheric scintillation. Values greater than 1 represent the occurrence of extremely severe ionospheric scintillation. The scintillation levels are normally classified into three categories [32–34]: strong if $0.6 \leq S4$, moderate if $0.2 < S4 \leq 0.6$, and weak if $0 < S4 \leq 0.2$. Due to the occasional contingency, irregularity and certain seasonal characteristics of ionospheric scintillation, there are far fewer strong scintillation events than weak and moderate scintillation events [35]. Moreover, most of the scintillation activity over the course of a day occurs from sunset to dawn [36]. In other words, the proportion of strong scintillation events in with respect to scintillation events of all intensities is relatively low, causing imbalance [16,20,23].

Thus, the imbalance exists naturally in ionospheric scintillation detection and monitoring. However, such a phenomenon has not been taken into consideration for the detection of interference in space weather monitoring.

2.2. Data Collection System

The data used in this paper were collected at the low-latitude site of São José dos Campos, Brazil (23.2S, 45.9W), from 2013 to 2015 during the last peak of the solar cycle, where the phenomenon of ionospheric scintillation was extremely active due to its geographical location close to both the South Atlantic Magnetic Anomaly (SAMA) and the EIA [37]. Months of GPS data from 2013 to 2015 were recorded, with strong ionospheric scintillation detected during each hour of monitoring. Aiming at studying the natural, occurrent and unpredictable phenomenon, a Septentrio PolaRx ionospheric scintillation monitoring (ISM) receiver was utilized to monitor the scintillation activity. Figure 2 shows the framework of the data collection system used to monitor ionospheric scintillation events. The wideband antenna was split into several ports that were connected to a commercial ISM receiver, Septentrio PolaRx, and SDR-based RF front ends, respectively. Scintillation-related measurements including channel correlation values were continuously collected by the ISM receiver, and relevant scintillation indices and event indicators were calculated simultaneously [10]. Furthermore, the indicators were to be compared with the threshold values preset before data collection to trigger the data server to record raw IF samples generated by the SDR front ends [38,39]. It should be mentioned that only data affected by the natural ionospheric scintillation phenomenon were to be recorded, to save a large amount of memory. These data could be utilized in research on optimal algorithms with respect to the ISM receiver, as well as providing a database for the analysis of strong scintillation characteristics. It is worth mentioning that, in the process of the acquisition and tracking, the DLL and PLL pull-in noise bandwidths were set as 2 Hz and 25 Hz, respectively, while the pull-in time was set as 500 ms. Once the signals were locked and the tracking loop was kept stable, the DLL and PLL pull-in noise bandwidths were reset to 1 Hz and 10 Hz, respectively, to reduce the influence of noise and other sources of interference.

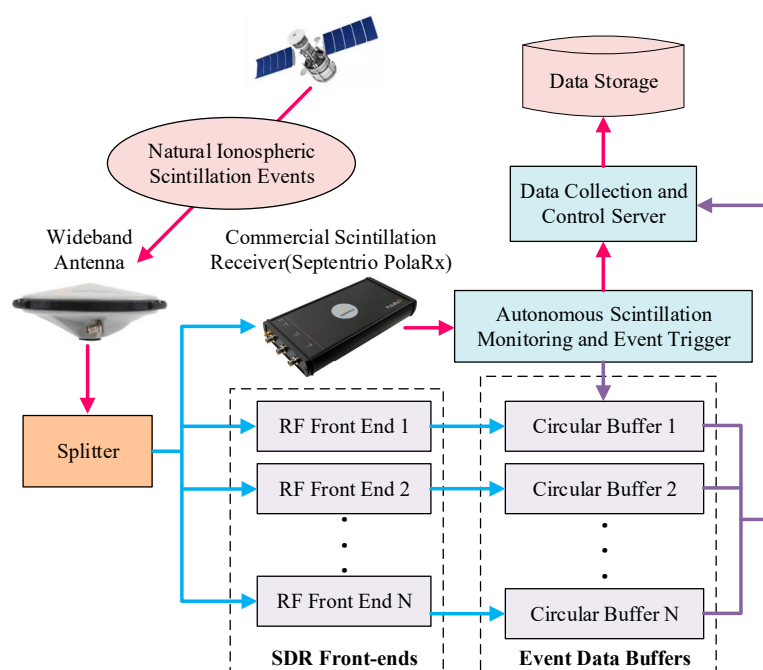


Figure 2. Architecture of the scintillation event-driven data collection system developed in the equatorial region. The commercial ISM Septentrio PolaRx is used to collect relevant navigation data and trigger SDR-based RF front ends when natural scintillation events occur.

A dataset of GPS L1 C/A signals was selected for the study. It was recorded from the GTEC free front ends, which were configured to collect zero-frequency data with 8-bit-resolution samples at a 20 MHz complex sampling rate. The collection period lasted from March 2013 to February 2015, covering the latest peak period for solar activity. Furthermore, these available data were processed by the SDR receiver advanced by a combination coherent/con-coherent integration acquisition algorithm to enhance the acquisition and tracking performance under the strong ionospheric scintillation environment.

Overall, 45 segments of one-hour data were used in the following research, with two or three satellites' data selected in each segment. The I/Q values were sampled at 1000 Hz, while the sample rate was set to 1 Hz according to the size of the shifting window. Based on three levels of scintillation intensity, Table 1 presents the information of the training dataset and testing dataset used in the following research.

Table 1. The distribution information of training dataset and testing dataset on three classes.

Dataset	Weak (Class 0)	Moderate (Class 1)	Strong (Class 2)	Total	Ratio
Training	155,892	136,343	38,515	349,559	4.05:3.54:1
Testing	18,680	14,496	5665	38,841	3.30:2.56:1

2.3. Feature Extraction

The features play a decisive role in the classification model. There are different sets of features producing different results for the detection accuracies, while the set of signal-based features as well as corresponding combinations are proved to be optimal with the location-independent technique [16]. The raw GNSS signal measurements in phase (I) and quadrature-phase (Q) correlators were utilized as features, which were extracted from the software GNSS receiver at the end of the tracking process. In recent research [16], the feature set of $\{\langle I \rangle, \langle Q \rangle, \langle I^2 \rangle, \langle Q^2 \rangle, \langle SI \rangle, \langle SI^2 \rangle\}$ —namely, the average values of I and Q correlator outputs, I^2 , Q^2 , signal intensity SI and SI^2 —has been proved to show excellent performance for ionospheric scintillation detection by machine learning algorithms. Thus, based on the above feature set and non-Gaussian noise characteristics of ionospheric scintillation [40], the other four metrics of variances and covariance are also presented to further reflect the features.

$$\sigma_{SI} = \frac{1}{N-1} \sum_{n=1}^N (SI_n - \langle SI \rangle)^2 \quad (1)$$

$$\text{cov}(I^2, Q^2) = \frac{1}{N-1} \sum_{n=1}^N (I_n^2 - \langle I^2 \rangle)(Q_n^2 - \langle Q^2 \rangle) \quad (2)$$

$$\sigma_I = \frac{1}{N-1} \sum_{n=1}^N (I_n - \langle I \rangle)^2 \quad (3)$$

$$\sigma_Q = \frac{1}{N-1} \sum_{n=1}^N (Q_n - \langle Q \rangle)^2 \quad (4)$$

where σ_{SI} , σ_I and σ_Q are variance of SI , I and Q , respectively, and $\text{cov}(I^2, Q^2)$ is the covariance of I^2 and Q^2 . N is the number of samples used in one value. To remove the impact of thermal noise and reserve the characteristics of the scintillation phenomenon, the frequency of the initial observables I and Q was $f = 1000$ Hz. For each feature, the observation window was set as $T = 60$ s, meaning that there were $N = T \cdot f = 60000$ initial observables averaged in the observation period. The frequency of observation was 1 Hz, which was in line with the shifting window of 1 s. Based on that, the features used in this work are described as follows.

$$X = \{\langle I \rangle, \langle Q \rangle, \langle I^2 \rangle, \langle Q^2 \rangle, \langle SI \rangle, \langle SI^2 \rangle, \sigma_{SI}, \text{cov}(I^2, Q^2), \sigma_I, \sigma_Q\} \quad (5)$$

The scintillation phenomenon may persist for several hours or longer, meaning that the data will be affected by the scintillation for a significant period of time rather than several minutes or less. The scintillation events were manually marked as $L = \{0, 1, 2\}$ based on the visual inspection of thresholds, divided into $0 < S_4 \leq 0.2$, $0.2 < S_4 \leq 0.6$ and $0.6 \leq S_4$, representing weak, moderate and strong scintillation events marked as 0, 1 and 2, respectively. Meanwhile, the corresponding carrier-to-noise density power ratio (C/N_0) was also taken into consideration in the manual detection process. This manual labelled approach combined with personal knowledge and experience can reserve the transient phases of the events and reduce the missed detections, significantly enhancing the labelling accuracy and detection performance.

3. Methodology

The goal of this approach was to propose an improved eXtreme Gradient Boosting (XGBoost) algorithm combined with the synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN), comparing the performance with that of the decision tree and random forest algorithms, which have been successfully used in recent research [16].

3.1. XGBoost Algorithm

The XGBoost algorithm was proposed by Chen Tianqi in 2016, presenting low computational complexity, a fast running speed and high accuracy [41]. As it is an inefficient ensemble learning algorithm, the boosting is aimed at transforming a weak classifier into a strong classifier to achieve good accuracy. Moreover, the gradient boosting attempts to improve robustness by making the algorithm's loss function drop along its gradient direction in the iteration process. Additionally, as a fast implementation of the gradient boosting algorithm, XGBoost can make full use of multi-core CPUs for parallel computation and improve the accuracy, significantly reducing the computational loads and enhancing the accuracy compared with other widely used algorithms such as the decision tree and random forest.

As the basis function of XGBoost, decision tree-based solutions for classification tasks have been used successfully in various fields, and corresponding advanced algorithms based on that are gradually being proposed to enhance the classification performance in terms of accuracy, precision, generalization and computing efficiency. There are three typically used decision-tree algorithms based on information theory: ID3, C4.5 and classification and regression tree (CART); the classification regulation of CART is derived from the Gini index, and CART is used more frequently than the two other algorithms. It not only assigns categories to leaf nodes, but also considers the possibility of all attributes being selected as leaf nodes. When all the samples in the node belong to the same attribution or the depth of the decision tree reaches a preset threshold, the tree construction stops. Due to the overfitting phenomenon for the former condition, it is vital to set the max depth value before training. Moreover, the random forest consists of multiple structurally similar decision trees for determining the dataset together to prevent overfitting and reduce the variance of an estimate. The training set is sampled to obtain multiple subsets, and then, several decision trees with the same structure are constructed, each of which is trained separately by using the constructed training subsets. The classification result depends on votes from all the decision trees trained on the testing set. More details can be found in the literature [16].

Consisting of different decision trees, the core of the XGBoost algorithm is learning a new function $f(x)$ to fit the last predicted residuals, adding decision trees continually and continuing to split features to grow a tree before meeting the growth conditions. After completing the training process and obtaining k trees, we needed to predict the novel data. According to the characteristics of samples, they will be classified to a leaf node corresponding to a score in each tree, and the sum of the scores for each tree to be regarded as the predicted value of the sample. Obviously, the goal of the algorithm is to make the

predicted value of all the trees \hat{y}_i as close as possible to the corresponding true value y_i , with as much generalization capability as possible. Based on the current tree, an additional tree is added to fit the residual between the predicted result of the previous trees and corresponding true value. To choose the next tree to be added, we introduced the objective function combined with the loss function and regular function, as shown in (6).

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + cons \quad (6)$$

where i represents the i th sample, n is the number of samples, y_i is the true score of the current tree, and $\hat{y}_i^{(t-1)}$ refers to the predicted value of $t-1$ trees. f_t is a new function of the current tree; $\Omega(f_t)$ is the corresponding regularization item. Additionally, the constant value $cons$ can be ignored without affecting the following objective process. The objective function can be approximated as follows after using Taylor expansion:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + cons \\ &= \sum_{i=1}^n \left[g_i \omega_j(x_i) + \frac{1}{2} h_i \omega_j^2(x_i) \right] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^T h_i \omega_j^2 + C \end{aligned} \quad (7)$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$; $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})$, T is the number of the leaf nodes, and ω_j is the weight of the j th leaf node. γ and λ are used to control the complexity of the trees. Defining $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$, the optimal solution $\omega_j^* = -\frac{G_j}{H_j + \lambda}$; then, the optimal objective function without a constant value C can be obtained as follows.

$$Obj^{(t)} = \sum_{j=1}^T \left[G_j \omega_j + \frac{1}{2} H_j \omega_j^2 \right] + \gamma T = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (8)$$

Furthermore, the greedy algorithm is utilized to enumerate all the different tree structures to find the optimal splitting node, achieving the maximum gain of the objective function after splitting, as shown in (9).

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (9)$$

where G_L and G_R are the first-order gradient values of the leaf nodes on left and right after splitting. Similarly, H_L and H_R are corresponding second gradient values. A brief schematic diagram of the XGBoost algorithm is shown in Figure 3.

There are some preset parameters used to control the tree building, such as γ , λ , the maximum depth of the tree, the minimum weight of the child node, and the learning rate for each tree. The main purpose is to limit the complexity and weight of each tree, so that the overfitting problem can be mitigated or prevented.

3.2. SMOTE-ENN Resampling Technique

An imbalance in the dataset refers to the phenomenon of certain types of samples being lower in number than other types, and researchers always have more interest in the minority class samples [42]. For similar characteristics of data, many classification algorithms have difficulty in accurately expressing the inherent features because of the lack of information on the minority class samples, which causes the decision boundary to be greatly compressed in the classification system. Although the overall accuracy of the detection model is excellent, it is unable to effectively detect the minority target samples that ought to be detected. As shown in Table 2, the weak, moderate and strong scintillation

events are marked as 0, 1 and 2, respectively. When the training sample proportion ranges from 1:1:1 to 7:7:1 based on a constant 100,000 entries, the overall accuracy remains relatively stable and good, ranging from 90.18% to 93.32%. Meanwhile, Figure 4 shows that the recall value for the majority class 0 remains above 93%, while that for class 1 keeps increasing and reaches over 98%. However, there is a distinct decline for class 2 from 87% to 74%, with an approximate 13% decline, which means that an increasing number of strong scintillation events go undetected. This indicates the great importance of balancing the minority class data in improving the accuracy of strong scintillation detection.

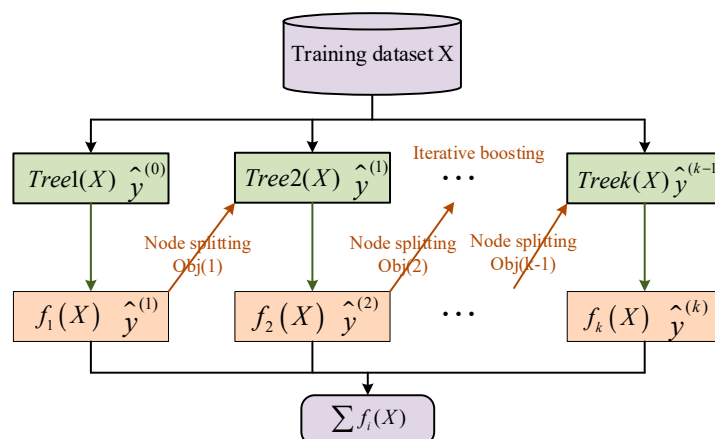


Figure 3. The schematic diagram of XGBoost algorithm.

Table 2. The overall accuracy values for different sample proportions for classes 0/1/2.

Sample Proportion (Class 0/1/2)	1:1:1	2:2:1	3:3:1	4:4:1	5:5:1	6:6:1	7:7:1
Overall Accuracy (%)	90.18	91.47	93.32	93.13	93.31	92.65	92.61

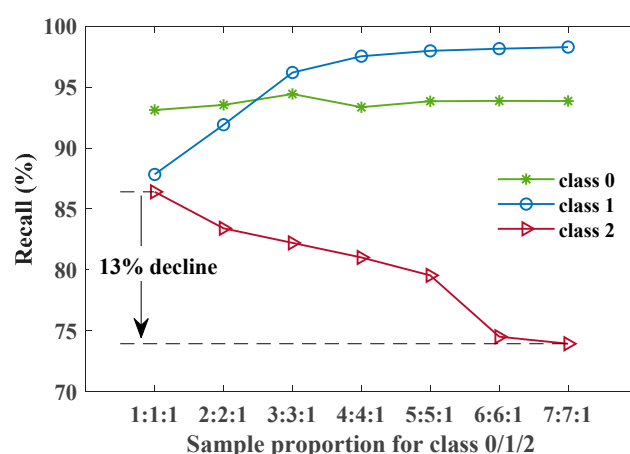


Figure 4. Influence of imbalanced datasets on performance of XGBoost model. These seven training datasets with 100,000 samples were selected randomly in the overall training dataset listed in Table 1, according to different sample ratios for classes 0/1/2. The testing dataset for each detection model trained by the above training dataset is the same as that mentioned in Table 1.

The purpose of resampling in this study was to balance the ionospheric scintillation events of different intensities, enhancing the detection accuracy for strong scintillation events. The normally used undersampling methods include random undersampling of the majority class, edited nearest neighbor (ENN) [43], Tomeklink [44], condensed nearest

neighbor (CNN) [45], and neighborhood cleaning (NCL) [46], while the typical oversampling methods include random oversampling of the minority class, the synthetic minority oversampling technique (SMOTE) [47], and Borderline-SMOTE (BSM) [48]. However, the former methods miss sample information through reducing the majority samples, and the latter methods add minority samples, causing a constant increase in useless information and an overfitting problem. The combinational algorithm SMOTE-ENN preserves the features of majority samples and increases the characteristics of minority samples, resulting in excellent classification performance for imbalanced datasets [49].

The basic idea of the SMOTE method is to carry out linear interpolation between neighboring minority class samples to synthesize new minority class samples, solving the problem of significant data overlap compared with random oversampling [47]. Further details can be described as follows: for each minority class sample $x_m \in X$ ($m = 1, \dots, M$), find the k nearest neighboring samples of the same class K , and then, choose n samples in K according to the sampling rate n , mark them as $y_{m,i} \in Y_m$ ($i = 1, \dots, n$, $m = 1, \dots, M$), and finally achieve random linear interpolation on the lines between x_m and $y_{m,1}, \dots, y_{m,n}$, respectively. The new built sample can be described as (10).

$$X_{new,m} = x_m + r * (Y_m - x_m), m = 1, \dots, M \quad (10)$$

where r is the random coefficient ranging from 0 to 1, and $X_{new,m}$ is the vector including n new samples built by x_m . In total, there are $M * n$ new samples.

Focusing on majority class samples, the ENN algorithm deletes the sample if there are two or more in the nearest three neighboring samples different from it [43]. However, the majority samples are near each other, which causes limited sample removal.

The SMOTE-ENN method achieves oversampling on the minority class samples, firstly using SMOTE, and then finishes undersampling on the majority class samples to coalesce their advantages. Thus, the combination of SMOTE and ENN is utilized to balance data with few strong ionospheric scintillation events; its performance was evaluated and compared with that of other methods, such as SMOTE and ENN. Figure 5 shows the process of resampling, training and predicting, using the XGBoost algorithm improved by the SMOTE-ENN resampling technique.

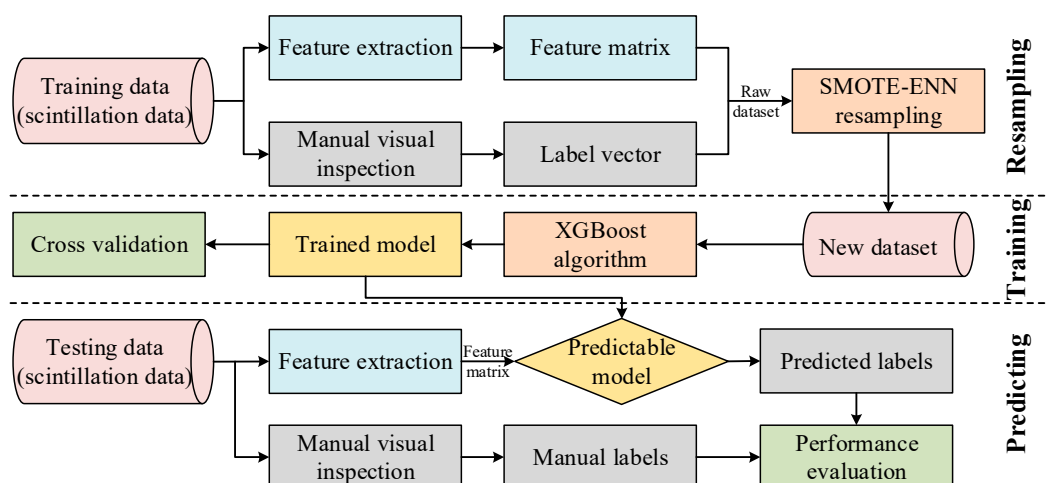


Figure 5. The process flow of XGBoost detection model improved by SMOTE-ENN resampling technique.

4. Results

A series of experiments were conducted to evaluate the performance of the proposed method and compare it with that of the decision tree and random forest algorithms, which have shown high validation accuracy in previous research [16]. The evaluation criterion with which the results were quantitatively analyzed is introduced. Based on the validation

phase, the overall accuracy of three different machine learning methods in the condition of different tree depths was compared, determining the complexity, computational efficiency and performance of these models to a large extent. Next, different resampling methods were implemented on the training dataset, comparing their performance in improving the detection of the minority class. Finally, experiments on training sets with different ratios of the three classes were carried out to evaluate the influence of imbalanced data on scintillation detection, while previous similar experiments were conducted on testing sets.

4.1. Evaluation Criterion

As one of the statistical tools most commonly used to evaluate the performance of detection algorithms, the confusion matrix, aimed at detection for three classes, was introduced. According to that, the corresponding metrics of overall accuracy, precision, recall and F1 score could be calculated to assess the performance of the detection models. Table 3 shows the example of the confusion matrix on three classes.

Table 3. The confusion matrix on three classes.

Class		Prediction		
		0	1	2
Truth	0	N_{00}	N_{01}	N_{02}
	1	N_{10}	N_{11}	N_{12}
	2	N_{20}	N_{21}	N_{22}

In Table 3, N_{ij} ($i, j = 0, 1, 2$) refers to the number of samples corresponding to the truth class i and prediction class j . According to that, the relative evaluation indicators on three classes can be described as follows:

$$accuracy = \frac{N_{00} + N_{11} + N_{22}}{N_{00} + N_{10} + N_{20} + N_{01} + N_{11} + N_{21} + N_{02} + N_{12} + N_{22}} \quad (11)$$

$$\begin{cases} precision_0 = \frac{N_{00}}{N_{00} + N_{10} + N_{20}} \\ precision_1 = \frac{N_{01}}{N_{01} + N_{11} + N_{21}} \\ precision_2 = \frac{N_{02}}{N_{02} + N_{12} + N_{22}} \end{cases} \quad (12)$$

$$\begin{cases} recall_0 = \frac{N_{00}}{N_{00} + N_{01} + N_{02}} \\ recall_1 = \frac{N_{10}}{N_{10} + N_{11} + N_{12}} \\ recall_2 = \frac{N_{20}}{N_{20} + N_{21} + N_{22}} \end{cases} \quad (13)$$

$$\begin{cases} F1-score_0 = \frac{2 * precision_0 * recall_0}{precision_0 + recall_0} \\ F1-score_1 = \frac{2 * precision_1 * recall_1}{precision_1 + recall_1} \\ F1-score_2 = \frac{2 * precision_2 * recall_2}{precision_2 + recall_2} \end{cases} \quad (14)$$

where the *precision* value represents the correct predicted positive ratio, and the *recall* value refers to the percentage of correctly predicted positive events among real positive events. Based on that, the weighted average of the precision and recall F1 score is defined as the F1 score.

4.2. Accuracy Evaluation on Cross Validation

As an important parameter for algorithms based on the decision tree, random forest and XGBoost, the tree depth determines the complexity and performance of models. Excessive depth might increase the validation accuracy, but can also lead to increased model complexity and greater computational loads, causing overfitting. Therefore, an appropriate tree depth is vital for the training model. Based on cross-validation, Figure 6 shows the mean accuracies across different tree depths and detection methods. Of the overall 349,559 points of the training set, 100,000 entries were randomly selected and used

in the validation experiment to reduce the running time. The accuracies of XGBoost are marked, ranging from 1 to 30, and compared with those of the decision tree and random forest, which presented high validation accuracies in previous research [16].

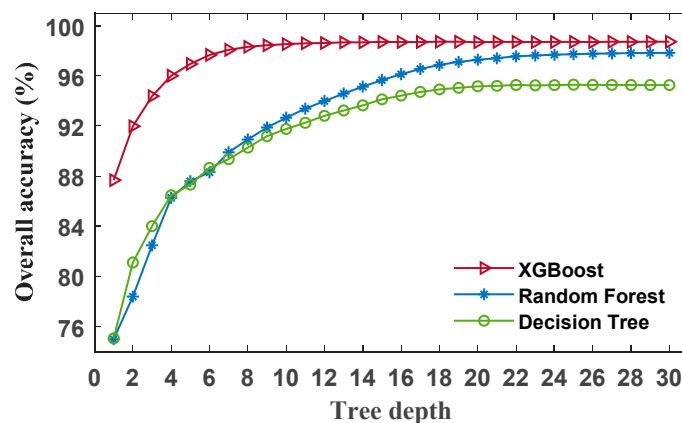


Figure 6. The mean accuracy of 10-fold cross validation based on different tree depths and detection methods. In the overall training dataset mentioned in Table 1, 100,000 samples were selected randomly and used in the validation process, and the remaining data were used as training samples.

Overall, the accuracy of XGBoost remained higher than that of the decision tree and random forest, which presented similar accuracy at different tree depths. Meanwhile, the high accuracy of XGBoost started to stabilize when the tree depth reached 6, while there was a significant advantage in accuracy compared with the other methods. The higher accuracy of XGBoost with a lower tree depth means a simpler trained model and less running time, compared with the random forest, which relies on a larger tree depth and is more computationally expensive. Even though the decision tree model shows a high training efficiency, the accuracy was significantly lower than that of XGBoost, even when the tree depth reached 20 or higher, causing overfitting on the validation dataset and affecting the detection performance for novel data. The XGBoost algorithm proved to be appropriate, with high accuracy and low computational expense. Other, more vital parameters were set to default because there were no huge training data or large feature classes.

Considering the running time and detection accuracy, the tree depth of XGBoost was set to 3 to reduce overfitting and improve the efficiency. The number of samples is also important for dealing with overfitting and underfitting problems. Based on the overall 349,559 points of the training set, 30% of the data were randomly reserved as an additional validation set, while 70% of the data were the training set, divided into 10, and the training size ranged from 0.1 to 1. As shown in Figure 7, the training accuracy across ten cross-validations and testing accuracy was compared at 10 different ratios for the training set. With an increase in training samples, the validation accuracy and testing accuracy gradually approached the same value and presented lower differences, which means that the model showed low bias and variance. Moreover, when the ratio of the training samples was set to 0.4 or higher, the accuracy variation decreased slightly and presented little fluctuation. To prevent overfitting or underfitting, the ratio was set to 0.4, meaning that approximately 100,000 points would be trained in following experiments.

4.3. Performance with Different Resampling Algorithms

Aiming at mitigating the impact of data imbalance on the detection of strong scintillation, a combination of oversampling and undersampling algorithms was proposed to achieve resampling. As shown in Table 4, comparisons were made between the SMOTE-ENN algorithm and other several single oversampling or undersampling methods, namely, the random oversampling, random undersampling, SMOTE, and ENN algorithms, on the basis of the raw data used in the training process. All of the comparison experiments were

based on 100,000 points of data randomly selected from the overall training dataset, and the ratio of weak, moderate and strong scintillation events was 4.05:3.54:1, in accord with those in the overall training dataset. The training process was achieved with the XGBoost algorithm with a tree depth of 3. The raw data were trained directly on the XGBoost model, while the other methods required resampling on the raw data and then training on the XGBoost model. All the trained models were tested on the 38,841 points of the dataset mentioned in Table 1 with a ratio of weak, moderate and strong scintillation events equal to 3.30:2.56:1, including 11 segments of one-hour data that were not involved in the training phase.

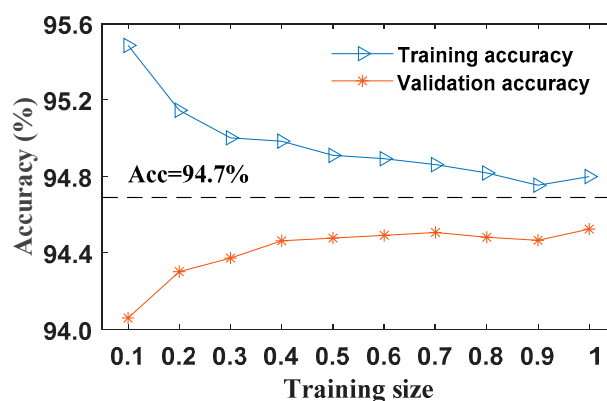


Figure 7. Training accuracy and testing accuracy of XGBoost versus the sample number used in the training set. In the overall training dataset mentioned in Table 1, 30% of the samples were randomly selected as the validation set, and the reserved samples were randomly selected as the training datasets according to different training sizes.

Table 4. Performance comparison for XGBoost algorithm improved by different resampling techniques based on testing results.

Resampling Technique		Raw Data	Random Oversampling	Random Undersampling	SMOTE	ENN	SMOTE-ENN
Precision (%)	0	99.38	99.51	99.64	99.77	99.47	99.71
	1	82.45	83.07	85.00	85.05	86.67	87.96
	2	96.02	95.14	94.83	95.19	95.77	95.00
Recall (%)	0	87.91	87.55	89.43	89.31	92.13	92.16
	1	97.89	97.64	97.65	97.93	97.85	97.77
	2	86.55	90.15	90.75	91.17	87.47	91.60
F1 score (%)	0	93.29	93.15	94.26	94.25	95.66	95.79
	1	89.51	89.77	90.89	91.04	91.93	92.61
	2	91.01	92.58	92.75	93.14	91.43	93.27
Accuracy (%)		91.43	91.69	92.69	92.80	93.58	94.17

Based on the SMOTE-ENN resampling technique, Table 5 compares the experimental data for five detection algorithms. The decision tree and random forest, each of which presented excellent performance in the binary classification for ionospheric scintillation detection in recent research [16], were tested for comparison. Considering the good performance of the SVM in similar detection tasks [19–23], tests for such a method were performed and analyzed. The CNN model was also employed in the comparison due to its effectiveness in a wide range of classification problems [25–28]. However, the SMOTE-ENN resampling technique is meaningless in regard to improving CNN performance, as its space structure becomes disordered after the resampling process. Thus, the CNN test was carried out on raw data only, while the same tests on the SVM accord with those on the other three methods. Due to the mentioned problem of the CNN and excessive computational loads of the SVM, these two methods were not studied further. As shown in

Table 5, apart from the overall accuracy, the recall and F1 score for class 2 were recorded to illustrate the performance in detecting strong scintillation events. In addition, considering the unavailability of the resampling technique for the CNN, larger computational loads and running time for the SVM, and the overall better performance with XGBoost, further tests with both methods were not carried out. Further details can be found in the Discussion.

Table 5. Comparison of performance of different detection algorithms trained with raw dataset and SMOTE-ENN resampled dataset (apart from CNN algorithm), based on testing results.

Algorithm	Accuracy (%)			Recall for Class 2 (%)			F1 Score for Class 2 (%)		
	Raw Data	SMOTE-ENN	Improvement Ratio	Raw Data	SMOTE-ENN	Improvement Ratio	Raw Data	SMOTE-ENN	Improvement Ratio
CNN	89.86	-	-	73.63	-	-	82.15	-	-
SVM	93.39	93.52	0.13	90.49	95.06	4.57	89.46	89.07	-0.39
Decision tree	85.39	87.36	1.97	86.73	90.41	3.68	88.94	90.98	2.04
Random forest	86.95	88.92	1.97	85.44	89.97	4.53	90.43	92.91	2.48
XGBoost	91.43	94.17	2.74	86.55	91.60	5.05	91.04	93.27	2.23

When training on the raw data and resampled data, the accuracy of XGBoost was about 4% and 7% better than that of the decision tree and random forest, as shown in the left panel of Figure 8. The middle and right panels in Figure 8 also show higher recall and F1 scores for the XGBoost method, implying the superiority of XGBoost improved by the SMOTE-ENN technique.

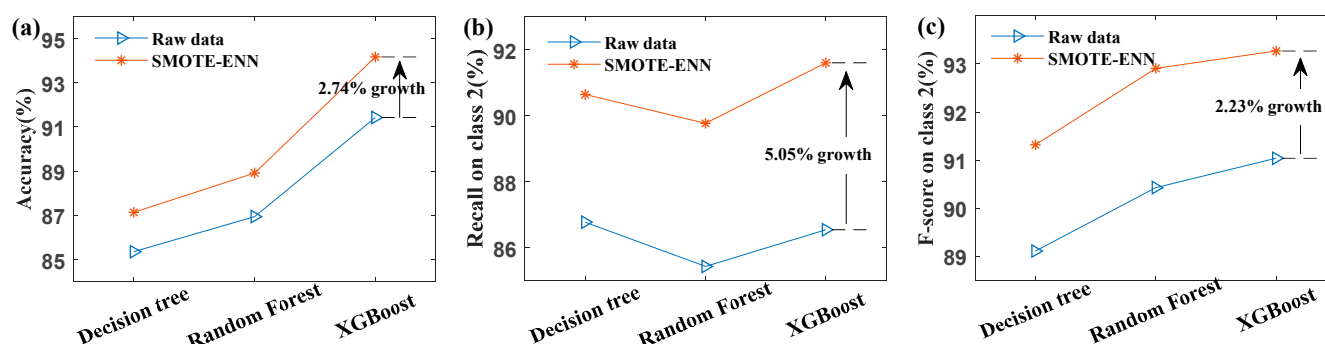


Figure 8. Comparison of detection algorithms in terms of overall accuracy, recall value for class 2 and F1 scores for class 2, respectively. From left to right: (a) the overall accuracy; (b) the recall value for class 2; (c) the F1 score for class 2. The blue polyline marked by triangles refers to the testing results based on raw training data, while the orange polyline marked by asterisks represents the testing results based on SMOTE-ENN resampled training data. The 100,000 training samples were selected in the overall training dataset listed in Table 1, while the testing dataset was that mentioned in Table 1.

4.4. Analysis on Imbalanced Training Datasets

To explore the detection performance of the XGBoost algorithm improved by the SMOTE-ENN resampling technique for strong scintillation events, a series of comparison experiments were performed based on different degrees of imbalance in the training and testing datasets, compared with the decision tree and random forest algorithm. From the training dataset with 349,559 samples, seven subsets of data were extracted. For each subset, 100,000 randomly selected samples were included, with the ratio of classes 0, 1 and 2 ranging from 1:1:1 to 7:7:1, respectively. All of the trained models were tested on the same dataset mentioned in Table 1.

According to the experimental results for the three detection algorithms and that improved by the SMOTE-ENN resampling technique based on these training subsets, Figure 9 shows the corresponding trends of each evaluation indicator with sample ratios of 1:1:1 to 7:7:1. Table 6 presents the improvement ratios in detail, based on a comparison between the detection results obtained with training on the resampled datasets and raw datasets.

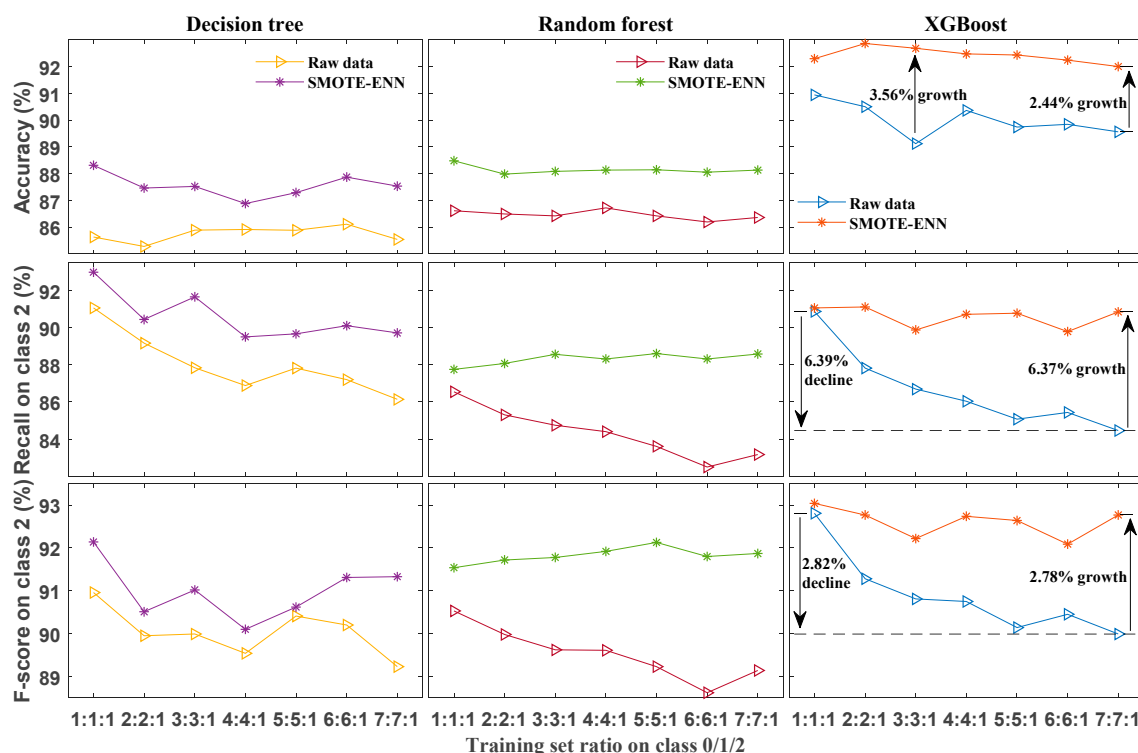


Figure 9. Testing results for detection algorithms regarding overall accuracy (top row), recall value for class 2 (middle row) and F1 score for class 2 (bottom row), with different ratios of classes 0/1/2 for the training sets. From left to right: testing results with decision tree (left column); testing results with random forest (middle column); testing results with XGBoost (right column). These seven training datasets with 100,000 samples were randomly selected from the overall training dataset listed in Table 1, according to different sample ratios of classes 0/1/2. The testing dataset for each detection model trained on the above training dataset is the same as that mentioned in Table 1.

Table 6. Improvement ratios for evaluation indicators with different training datasets with different detection algorithms trained on SMOTE-ENN resampled data, compared with raw training data.

Training Set		1	2	3	4	5	6	7
Accuracy (%)	Decision tree	2.68	2.18	1.63	0.97	1.41	1.76	1.99
	Random forest	1.87	1.49	1.66	1.41	1.72	1.86	1.77
	XGBoost	1.35	2.36	3.56	2.11	2.69	2.4	2.44
Recall for class 2 (%)	Decision tree	1.92	1.27	3.81	2.62	1.84	2.9	3.57
	Random forest	1.2	2.76	3.81	3.9	4.98	5.8	5.4
	XGBoost	0.19	3.28	3.18	4.67	5.69	4.34	6.37
F1 score for class 2 (%)	Decision tree	1.18	0.56	1.03	0.56	0.21	1.11	2.1
	Random forest	1.01	1.74	2.16	2.31	2.9	3.18	2.73
	XGBoost	0.23	1.49	1.41	1.99	2.5	1.64	2.78

4.5. Analysis on Imbalanced Testing Datasets

We also investigated the detection performance of the proposed method for strong scintillation events with different degrees of imbalance in the testing datasets. Table 7 lists the training datasets and testing datasets with different ratios of the three classes. The same 100,000 samples randomly selected in the overall training dataset were used in each model training process, while seven groups of testing datasets were designed with different ratios of classes, roughly ranging from 1:1:1 to 7:7:1. For each testing dataset with 31,779 samples, the dataset consisted of nine segments of one-hour observations consecutively, with 3531 samples included in each hour's data. Additionally, several such

segments could be included in several groups of testing sets, but each group of the testing set was not involved in the training dataset.

Table 7. The distribution information for testing dataset imbalanced with respect to three classes.

Dataset		Scintillation Intensity			Total	Ratio
		Weak (Class 0)	Moderate (Class 1)	Strong (Class 2)		
Training set	Raw data	46,610	42,514	10,876	100,000	4.05:3.54:1
	SMOTE-ENN	43,201	41,468	45,976	130,645	1.04:1:1.11
Testing set	1	11,945	9264	10,570	31,779	1.13:0.87:1
	2	11,983	14,290	5506	31,779	2.18:2.60:1
	3	13,984	13,052	4743	31,779	2.95:2.75:1
	4	14,124	14,595	3060	31,779	4.61:4.77:1
	5	14,124	15,102	2553	31,779	5.53:5.92:1
	6	16,268	13,140	2371	31,779	6.86:5.54:1
	7	14,124	15,728	1927	31,779	7.33:8.16:1

Table 8 records the improvement ratios for the evaluation indicators for different testing datasets with different detection algorithms trained on SMOTE-ENN resampled data, compared with raw training data, while Figure 10 shows the corresponding values of accuracy, recall and F1 scores for class 2.

Table 8. Improvement ratios for evaluation indicators for different testing datasets with different detection algorithms trained on SMOTE-ENN resampled data, compared with raw training data.

Testing Set		1	2	3	4	5	6	7
Accuracy (%)	Decision tree	2.58	1.57	2.57	0.47	1.18	0.44	1.06
	Random forest	2.45	2.07	2.34	0.65	1.24	0.36	0.45
	XGBoost	6.04	2.44	2.92	0.94	0.62	1.25	0.29
Recall for class 2 (%)	Decision tree	4.09	2.58	10.94	9.15	9.24	8.01	8.41
	Random forest	5.82	9.23	10.61	4.51	13.90	7.00	12.61
	XGBoost	10.59	12.60	17.69	13.98	16.17	19.40	14.06
F1 score for class 2 (%)	Decision tree	2.28	1.76	6.53	2.76	5.46	1.06	4.51
	Random forest	3.17	5.01	5.62	1.29	7.5	2.11	5.37
	XGBoost	6.00	6.64	9.54	7.15	8.04	10.61	6.61

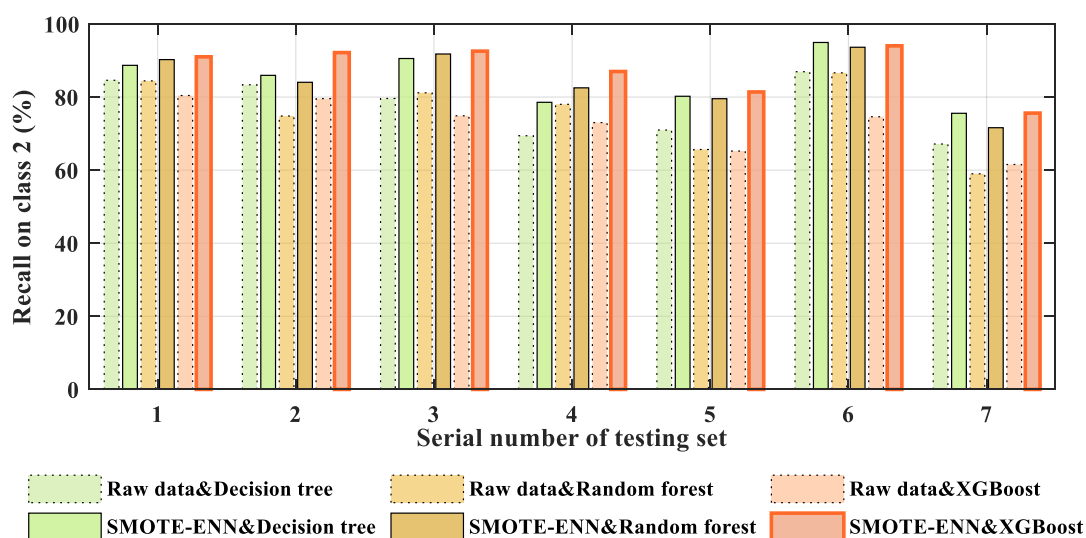


Figure 10. Cont.

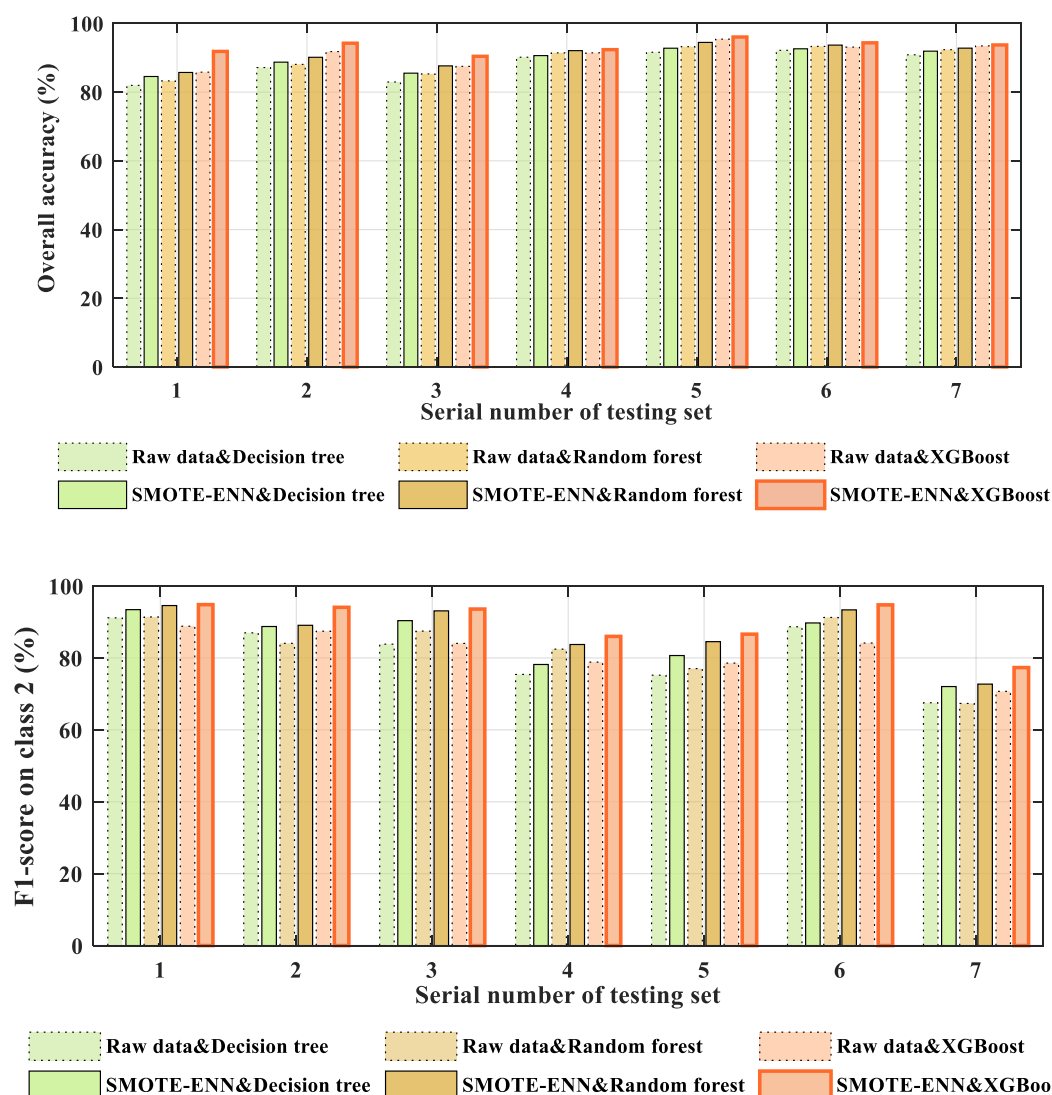


Figure 10. Testing results for detection algorithms regarding overall accuracy (top row), recall value for class 2 (middle row) and F1 score for class 2 (bottom row), with different ratios of classes 0/1/2 for testing datasets. Relevant information on training dataset and seven testing datasets is listed in Table 7. The green, yellow and orange bars refer to results for decision tree, random forest and XGBoost, respectively. The highlighted bold bar on the right of each group represents the result of the XGBoost method improved by the SMOTE-ENN technique, with better performance.

5. Discussion

All the experiments were carried out on real data collected in an equatorial region during solar maximum years, under various ionospheric scintillation intensities. Firstly, different resampling techniques were compared based on the XGBoost algorithm, and the integration of SMOTE and ENN produced better results regarding XGBoost's performance. As shown in Table 4, based on random oversampling, there was only a 3% increase in the recall value for the minority class 2 in that it just randomly took out samples in class 2 without any rule, and thus added in several invalid data. Meanwhile, there was little contribution to the other two classes. Different from oversampling, the method of random undersampling is aimed at randomly deleting samples of the majority class until the data of these classes are equal to those of the minority classes in number. Thus, much useful information is discarded, reducing the detection accuracy, presenting similar results to random oversampling. Although the resampling rule of SMOTE is based on the nearest neighboring samples and linear interpolation, it still aims at increasing the minority class

without significant improvement for the other two classes. As for the SMOTE, ENN only showed a 4% improvement on the recall value of the majority class 0, because the samples of the boundary around classes 0 and 1 were deleted. Meanwhile, the overall samples were still imbalanced, resulting in ineffective enhancement for class 2.

Differently, the SMOTE-ENN method firstly deals with the data of the minority class through the SMOTE oversampling technique and then carries out the ENN undersampling for the data of the majority class. It not only increases the data of the minority class but also deletes the multi-class samples on the boundary as much as possible, significantly improving the detection performance in various aspects. The results in Table 4 show that there was an apparent improvement in performance for all three classes, with an approximately 3% increase in overall accuracy. As for the value of recall, it showed a 5% enhancement for class 2, corresponding to strong scintillation events, and 4% increase for class 0, corresponding to weak scintillation events, meaning that more strong and weak scintillation events that may be easily missed can be detected correctly. Thereby, the precision for class 1, corresponding to moderate scintillation events, increased by 6%, with more events of classes 0 and 2 being correctly detected. Meanwhile, the F1 scores for the three classes increased by 3%, 3% and 2%, respectively, demonstrating an improvement in both precision and recall. Especially, the integrated technique outperforms ENN by 4.13% in recall for class 2, while showing a 1.37% enhancement in overall accuracy compared with SMOTE. Compared with the other single resampling methods, the SMOTE-ENN proved to be effective in dealing with imbalanced data with insufficient strong scintillation events. Based on that, brief comparisons of five methods were drawn. As shown in Table 5, XGBoost outperformed the CNN by 1.57%, 12.92% and 8.99% in terms of the overall accuracy, recall and F1 score for class 2 when trained and tested on raw data, showing significant superiority in strong scintillation detection. Overall, all four detection algorithms were enhanced when improved by SMOTE-ENN. Moreover, for resampled data, XGBoost outperformed the SVM by 0.65% in terms of overall accuracy. The recall for the SVM was 3.46% higher than that for XGBoost, but the F1 score for the SVM was 4.20% lower. This indicates that the SVM missed many strong scintillation events, with a precision of only 83.79% for class 2. More importantly, larger computational loads and longer running times are needed for SVM detection than detection with the three other machine learning methods. Thus, both methods will not be further discussed due to the unavailability of the CNN and large running loads for the SVM.

For the different imbalanced training sets, the accuracy of the improved XGBoost was about 4% to 5% higher than that of the decision tree and random forest, which is clear from Figure 9. Although the recall values of the decision tree were about 2% higher than those of XGBoost based on the training sets of 1, 3 and 6, the F1 scores for XGBoost were about 2% higher than those for the decision tree and random forest. As shown in the right panels of Figure 9, there was a 6.39% decline in the recall value, from 90.86% to 84.47%, and 2.82% decline in the F1 score, from 92.81% to 89.99%, implying more missed strong scintillation events. However, the XGBoost model, trained on the SMOTE-ENN resampled datasets, remained relatively stable, with recall values of about 90% to 91% and F1 scores of 92% to 93%, indicating more enhancement for more severely imbalanced training samples. Meanwhile, there was a 6.37% improvement in the recall value, from 84.47% to 90.84%, and 2.78% improvement in the F1 score, from 89.99% to 92.77%, with the resampled training data with a sample ratio of 7:7:1. Additionally, there was also improvement in the other two classes in the precision, recall and F1 score, as well as the overall accuracy, which increased by 3% to over 92% compared to that with the raw data. These illustrate the effectiveness of the proposed XGBoost algorithm improved by the SMOTE-ENN resampling technique for imbalanced training data in improving the detection performance for strong scintillation events.

Considering various natural phenomena of different ionospheric scintillation events, seven groups of testing datasets with different degrees of imbalance were established, and each dataset included nine segments of real one-hour data. The results significantly illus-

trate the excellent performance of the improved XGBoost in strong scintillation detection under different scintillation conditions. It can be intuitively observed from Figure 10 that the accuracy of the improved XGBoost was higher than that of the other methods based on each testing dataset, while almost all the values of recall and F1 score were similar. The overall testing accuracy of the XGBoost model trained on the resampled dataset varied from 90.42% to 96.04%, higher than that of the XGBoost model trained on the raw dataset, ranging from 85.81% to 95.42%. Table 8 illustrates the improvement ratios for the evaluation indicators for the different testing sets with different detection algorithms trained on the resampled dataset, compared with the raw dataset. For these testing datasets, the results show an increase of at least 10% in the recall values. Especially, for the sixth group of the testing dataset, it increased by 19.40%, from 74.61% to 94.01%. Meanwhile, the F1 score significantly increased by more than 6.00%, ranging from 77.31% to 94.82%, for these seven testing datasets, while there was a significant increase of 10.61% from 84.14% to 94.75% for the sixth testing dataset. The results also present an improvement or maintenance of the corresponding recall, precision and F1 scores for the other two classes. These results indicate that it is valuable to enhance the detection accuracy for strong scintillation events with different degrees of imbalance in the testing data with the method of resampling the imbalanced training data by SMOTE-ENN before training the XGBoost model.

6. Conclusions

Severe ionospheric scintillation is an adverse factor influencing the amplitude and carrier phase of a GNSS signal; its detection is a prerequisite in the design of an advanced receiver with greater accuracy, reliability and efficiency. Nevertheless, the natural appearance of strong ionospheric scintillation occurs incidentally compared to that of weak/moderate scintillation. The imbalance may prove a challenge in achieving higher detection accuracy for strong scintillation events. As a strategy for detecting the severe ionospheric scintillation events, the eXtreme Gradient Boosting (XGBoost) algorithm improved by the synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) resampling technique was developed as follows:

- (1) The machine learning method of XGBoost was proposed to improve the overall detection accuracy. According to 10 cross-validations, the accuracy was better than that of the decision tree and random forest. Meanwhile, XGBoost demonstrated sufficient validation accuracy when the tree depth was set to a small value, which not only significantly simplified the model complexity, but also effectively alleviated the overfitting problem.
- (2) Aiming at dealing with imbalance, different resampling techniques were compared based on the XGBoost detection model. SMOTE-ENN outperformed the other techniques on the whole. Moreover, similar improvements were observed for the decision tree and random forest detection models, after the SMOTE-ENN resampling technique, while the improved XGBoost performed better than the other methods.
- (3) As for training datasets with different degrees of imbalance in classes 0/1/2, with ratios ranging from 1:1:1 to 7:7:1, different detection models and corresponding models improved by SMOTE-ENN were trained and then tested on the same novel dataset. The results showed overall enhancements for the improved detection methods compared to the corresponding raw methods, among which the improved XGBoost method showed the best performance.
- (4) The improved methods were tested with different degrees of imbalance in real data to evaluate the performance of the improved XGBoost. The results show distinct enhancements in overall accuracy, recall for class 2 and F1 scores for class 2, proving significant improvements in detecting severe scintillation events as well as reducing the problem of missing important events.

Consequently, the performance of XGBoost improved by SMOTE-ENN was examined in comparative tests and under various conditions. The significance of these results lies in dealing with the problems of the natural and incidental appearance of strong scintillation

events, which may cause imbalance and especially affect the detection accuracy for strong scintillation events. This work would be of general interest for researchers in the fields of detecting interference in satellite signals (e.g., ionospheric scintillation, solar radio burst, and spoofing); the design of advanced receivers with greater accuracy, reliability and efficiency; and the atmospheric layer and space weather.

Author Contributions: Conceptualization, M.L.; methodology, M.L.; software, X.Z.; validation, X.T. and G.T.; formal analysis, M.L.; writing—original draft preparation, M.L.; writing—review and editing, T.H.; supervision, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Plan of China, grant number 2018YFB0505103, and the National Natural Science Foundation of China, grant number 61873064.

Data Availability Statement: Relevant data are available on request from the corresponding author. Please contact Xuefen Zhu (zhuxuefen@seu.edu.cn).

Acknowledgments: The authors would like to thank three anonymous reviewers for their valuable and insightful comments.

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

References

1. Cai, X.; Burns, A.G.; Wang, W.; Qian, L.; Liu, J.; Solomon, S.C.; Eastes, R.W.; Daniell, R.E.; Martinis, C.R.; McClintock, W.E. Observation of Postsunset OI 135.6 nm radiance enhancement over South America by the GOLD mission. *J. Geophys. Res. Space Phys.* **2020**, *126*, 2020JA028108.
2. Karan, D.K.; Daniell, R.E.; England, S.L.; Martinis, C.R.; Eastes, R.W.; Burns, A.G.; McClintock, W.E. Early morning equatorial ionization anomaly from GOLD observations. *J. Geophys. Res. Space Phys.* **2020**, *125*, 2019JA027487.
3. Karan, D.K.; Daniell, R.E.; England, S.L.; Martinis, C.R.; Eastes, R.W.; Burns, A.G.; McClintock, W.E. First zonal drift velocity measurement of Equatorial Plasma Bubbles (EPBs) from a geostationary orbit using GOLD data. *J. Geophys. Res. Space Phys.* **2020**, *125*, e2020JA028173. [[CrossRef](#)]
4. Martinis, C.; Daniell, R.; Eastes, R.; Norrell, J.; Smith, J.; Klenzing, J.; Solomon, S.; Burns, A. Longitudinal variation of post-sunset plasma depletions from the Global-scale Observations of the Limb and Disk (GOLD) mission. *J. Geophys. Res. Space Phys.* **2020**, *126*, 2020JA028510.
5. Liu, J.; Wang, W.; Burns, A.; Solomon, S.C.; Zhang, S.; Zhang, Y.; Huang, C. Relative importance of horizontal and vertical transports to the formation of ionospheric storm-enhanced density and polar tongue of ionization. *J. Geophys. Res. Space Phys.* **2016**, *121*, 8121–8133. [[CrossRef](#)]
6. Basu, S.; MacKenzie, E.; Basu, S. Ionospheric constraints on VHF/UHF communications links during solar maximum and minimum periods. *Radio Sci.* **1988**, *23*, 363–378. [[CrossRef](#)]
7. Sreeja, V.; Aquino, M.; Elmas, Z.G. Impact of ionospheric scintillation on GNSS receiver tracking performance over Latin America: Introducing the concept of tracking jitter variance maps. *Space Weather* **2011**, *9*, S10002. [[CrossRef](#)]
8. Aarons, J.; Whitney, H.E.; Allen, R.S. Global morphology of ionospheric scintillations. *Proc. IEEE* **1971**, *59*, 159–172. [[CrossRef](#)]
9. Knepp, D.L. Radar measurement of ionospheric scintillation in the polar region. *Radio Sci.* **2015**, *50*, 968–982. [[CrossRef](#)]
10. Jiao, Y.; Morton, Y.T. Comparison of the effect of high-latitude and equatorial ionospheric scintillation on GPS signals during the maximum of solar cycle 24. *Radio Sci.* **2015**, *50*, 886–903. [[CrossRef](#)]
11. Prasad, S.N.V.S.; Rama Rao, P.V.S.; Prasad, D.S.V.V.D.; Venkatesh, K.; Niranjana, K. Morphological studies on ionospheric VHF scintillations over an Indian low latitude station during a solar cycle period (2001–2010). *Adv. Space Res.* **2012**, *50*, 56–69. [[CrossRef](#)]
12. Banville, S.; Langley, R.B.; Saito, S.; Yoshihara, T. Handling cycle slips in GPS data during ionospheric plasma bubble events. *Radio Sci.* **2010**, *45*, 1–14. [[CrossRef](#)]
13. Ji, S.; Chen, W.; Weng, D.; Wang, Z.; Ding, X. A study on cycle slip detection and correction in case of ionospheric scintillation. *Adv. Space Res.* **2013**, *51*, 742–753. [[CrossRef](#)]
14. Taylor, S.; Morton, Y.; Jiao, Y.; Triplett, J.; Pelgrum, W. An improved ionosphere scintillation event detection and automatic trigger for a GNSS data collection system. In Proceedings of the Institute of Navigation International Technical Meeting 2012, Newport Beach, CA, USA, 30 January–1 February 2012; pp. 1563–1569.
15. Dubey, S.; Wahi, R.; Gwal, A.K. Ionospheric effects on GPS positioning. *Adv. Space Res.* **2006**, *38*, 2478–2484. [[CrossRef](#)]
16. Linty, N.; Farasin, A.; Favenza, A.; Dovis, F. Detection of GNSS ionospheric scintillations based on machine learning decision tree. *IEEE Trans. Aerosp. Electron. Syst.* **2019**, *55*, 303–317. [[CrossRef](#)]

17. Mushini, S.C.; Jayachandran, P.T.; Langley, R.B.; MacDougall, J.W.; Pokhotelov, D. Improved amplitude- and phase-scintillation indices derived from wavelet detrended high-latitude GPS data. *GPS Solut.* **2012**, *16*, 363–373. [\[CrossRef\]](#)
18. Ouassou, M.; Kristiansen, O.; Gjevestad, J.G.O.; Jacobsen, K.S.; Andalsvik, Y.L. Estimation of scintillation indices: A novel approach based on local kernel regression methods. *Int. J. Navig. Obs.* **2016**. [\[CrossRef\]](#)
19. Jiao, Y.; Hall, J.; Morton, Y. Performance evaluations of an equatorial GPS amplitude scintillation detector using a machine learning algorithm. In Proceedings of the 29th International Technical Meeting of the Satellite Division of the Institute of Navigation, Portland, OR, USA, 12–16 September 2016; pp. 195–199.
20. Jiao, Y.; Hall, J.J.; Morton, Y.T. Automatic equatorial GPS amplitude scintillation detection using a machine learning algorithm. *IEEE Trans. Aerosp. Electron. Syst.* **2017**, *53*, 405–418. [\[CrossRef\]](#)
21. Jiao, Y.; Hall, J.; Morton, Y. Automatic GPS phase scintillation detector using a machine learning algorithm. In Proceedings of the 2017 International Technical Meeting of The Institute of Navigation, Monterey, CA, USA, 30 January–2 February 2017; pp. 1160–1172.
22. Jiao, Y.; Hall, J.J.; Morton, Y.T. Performance Evaluation of an automatic GPS ionospheric phase scintillation detector using a machine-learning algorithm. *Navig. J. Inst. Navig.* **2017**, *64*, 391–402. [\[CrossRef\]](#)
23. Lin, M.; Zhu, X.; Luo, Y.; Yang, F. Analysis of ionospheric scintillation detection based on machine learning. In Proceedings of the International Conference on Sensing, Measurement and Data Analytics in the Era of Artificial Intelligence, Xi'an, China, 15–17 October 2020; pp. 357–361.
24. Ludwig-Barbosa, V.; Sievert, T.; Carlström, A.; Pettersson, M.I.; Vu, V.T.; Rasch, J. Supervised detection of ionospheric scintillation in low-latitude radio occultation measurements. *Remote Sens.* **2021**, *13*, 1690. [\[CrossRef\]](#)
25. Ferre, R.M.; Fuente, A.D.L.; Lohan, E.S. Jammer classification in GNSS bands via machine learning algorithms. *Sensors* **2019**, *19*, 5–7.
26. Munin, E.; Blais, A.; Couellan, N. GNSS multipath detection using embedded deep CNN on Intel (R) Neural Compute Stick. In Proceedings of the 33rd International Technical Meeting of the Satellite Division of the Institute of Navigation, Online, 21–25 September 2020; pp. 2018–2029.
27. Suzuki, T.; Kusama, K.; Amano, Y. NLOS multipath detection using convolutional neural network. In Proceedings of the 33rd International Technical Meeting of the Satellite Division of the Institute of Navigation, Online, 21–25 September 2020; pp. 2989–3000.
28. Li, J.; Zhu, X.; Ouyang, M.; Li, W.; Chen, Z.; Dai, Z. Research on multi-peak detection of small delay spoofing signal. *IEEE Access* **2020**, *8*, 151777–151787. [\[CrossRef\]](#)
29. Franzese, G.; Linty, N.; Dovis, F. Semi-supervised GNSS scintillations detection based on deepinfomax. *Appl. Sci.* **2020**, *10*, 381. [\[CrossRef\]](#)
30. Dogo, E.M.; Nwulu, N.I.; Twala, B.; Aigbavboa, C. Accessing imbalance learning using dynamic selection approach in water quality anomaly detection. *Symmetry* **2021**, *13*, 818. [\[CrossRef\]](#)
31. Luo, X.; Lou, Y.; Xiao, Q.; Gu, S.; Chen, B.; Liu, Z. Investigation of ionospheric scintillation effects on BDS precise point positioning at low-latitude regions. *GPS Solut.* **2018**, *22*, 1–12. [\[CrossRef\]](#)
32. Marlia, D.; Wu, F.; Ekawati, S.; Anggarani, S.; Ahmed, W.A.; Nofri, E.; Byambakhuu, G. Ionospheric scintillation mapping at low latitude: Over Indonesia. In Proceedings of the International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2015; pp. 21–24.
33. Vadakke Veetil, S.; Aquino, M.; Marques, H.A.; Moraes, A. Mitigation of ionospheric scintillation effects on GNSS precise point positioning (PPP) at low latitudes. *J. Geod.* **2020**, *94*, 1–10. [\[CrossRef\]](#)
34. Jiao, Y.; Morton, Y.; Taylor, S.; Pelgrum, W. High latitude ionosphere scintillation characterization. In Proceedings of the Institute of Navigation International Technical Meeting, San Diego, CA, USA, 28–30 January 2013; pp. 579–584.
35. Merid, A.; Nigussie, M.; Ayele, A. Investigation of the characteristics of wavelike oscillations of post-sunset equatorial ionospheric irregularity by decomposing fluctuating TEC. *Adv. Space Res.* **2021**, *67*, 1210–1221. [\[CrossRef\]](#)
36. Guo, R. Statistical Studies of Radio Wave Amplitude and Phase Scintillation in the Ionosphere. Master's Thesis, Wuhan University, Wuhan, China, 2019.
37. Spogli, L.; Alfonsi, L.; Romano, V.; De Franceschi, G.; Joao Francisco, G.M.; Hirokazu Shimabukuro, M.; Bougard, B.; Aquino, M. Assessing the GNSS scintillation climate over Brazil under increasing solar activity. *J. Atmos. Solar-Terr. Phys.* **2013**, *105*–106, 199–206. [\[CrossRef\]](#)
38. Taylor, S.; Morton, Y.; Marcus, R.; Bourne, H.; Pelgrum, W.; Van Dierendonck, A.J. Ionospheric scintillation receivers performances based on high latitude experiments. In Proceedings of the Institute of Navigation Pacific Positioning, Navigation and Timing Meeting, Honolulu, HI, USA, 22–25 April 2013; pp. 743–751.
39. Xu, D.; Morton, Y.; Akos, D.; Walter, T. GPS multi-frequency carrier phase characterization during strong equatorial ionospheric scintillation. In Proceedings of the 28th International Technical Meeting of the Satellite Division of the Institute of Navigation, Tampa, FL, USA, 14–18 September 2015; pp. 3787–3796.
40. Sun, P. Study on Key Techniques of the Satellite Navigation Signals Carrier Tracking in the Presence of Ionospheric Scintillation. Ph.D. Thesis, Graduate School of National University of Defense Technology, Changsha, China, March 2017.
41. 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.
42. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.

-
43. Wilson, D.L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.* **1972**, *2*, 408–421. [[CrossRef](#)]
 44. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst. Man Cybern.* **1976**, *6*, 769–772.
 45. Hart, P.E. The condensed nearest neighbor rule. *IEEE Trans. Inf. Theory* **1967**, *14*, 515–516. [[CrossRef](#)]
 46. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe, Cascais, Portugal, 1–4 July 2001; pp. 63–66.
 47. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique nitesh. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
 48. Han, H.; Wang, W.; Mao, B. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the Advances in Intelligent Computing—International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; pp. 878–887.
 49. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [[CrossRef](#)]