universe

MDPI

Article

Random Forest Classification and Ionospheric Response to Solar Flares: Analysis and Validation

Filip Arnaut *, Aleksandra Kolarski and Vladimir A. Srećković

Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia; aleksandra.kolarski@ipb.ac.rs (A.K.); vlada@ipb.ac.rs (V.A.S.)
* Correspondence: filip.arnaut@ipb.ac.rs

Abstract: The process of manually checking, validating, and excluding data in an ionospheric very-low-frequency (VLF) analysis during extreme events is a labor-intensive and time-consuming task. However, this task can be automated through the utilization of machine learning (ML) classification techniques. This research paper employed the Random Forest (RF) classification algorithm to automatically classify the impact of solar flares on ionospheric VLF data and erroneous data points, such as instrumentation errors and noisy data. The data used for analysis were collected during September and October 2011, encompassing solar flare classes ranging from C2.5 to X2.1. The F1-score values obtained from the test dataset displayed values of 0.848; meanwhile, a more detailed analysis revealed that, due to the imbalanced distribution of the target class, the per-class F1-score indicated higher values for the normal data point class (0.69–0.97) compared to those of the anomalous data point class (0.31 to 0.71). Instances of successful and inadequate categorization were analyzed and presented visually. This research investigated the potential application of ML techniques in the automated identification and classification of erroneous VLF amplitude data points; however, the findings of this research hold promise for the detection of short-term ionospheric responses to, e.g., gamma ray bursts (GRBs), or in the analysis of pre-earthquake ionospheric anomalies.

Keywords: machine learning; solar flares; VLF ionospheric data; anomaly classification; anomaly detection; class imbalance

Citation: Arnaut, F.; Kolarski, A.; Srećković, V.A. Random Forest Classification and Ionospheric Response to Solar Flares: Analysis and Validation. *Universe* **2023**, *9*, 436. https://doi.org/10.3390/universe9100436

Academic Editor: Dmitrii Kolotkov

Received: 8 August 2023
Revised: 14 September 2023
Accepted: 28 September 2023
Published: 30 September 2023

1. Introduction

In recent years, the study of the terrestrial ionosphere in relation to various extreme phenomena has become a very interesting topic. Using various observation techniques, many scientists around the world have made significant progress in studying the imprints of extreme phenomena in the ionosphere [1,2]. As a result, the ionosphere is recognized as a useful "tool" for studying the disturbances caused by such phenomena. Modern communication systems, including satellite and navigation systems, as well as various radio signals, traverse the atmosphere and lower ionosphere. It is worth noting that these signals may encounter disturbances during periods of solar perturbations. X-ray solar flares can be classified based on their influence on VLF wave propagation via the Earth–ionosphere waveguide [3]. The significance of modeling the lower ionosphere cannot be overstated in relation to diverse technological, research, and industrial fields [4].

VLF ionospheric data modelling, like any other type of modelling, involves a preliminary data pre-processing phase. During this phase, data filtering and domain-specific transformations are usually applied, along with the elimination of erroneous data points. In the context of VLF ionospheric amplitude analysis, the methodology bears resemblance to the aforementioned approach. However, it diverges in the sense that the elimination of two distinct categories of erroneous data points is required. The first category pertains to instrumentation errors, wherein the VLF receiver produces flawed measurements. The second category involves the researcher's decision to either exclude or annotate the data

points affected by the solar flare influences on the VLF data. The process of manually eliminating and/or annotating erroneous data in VLF data analysis is subject to an additional consideration, namely, the high measurement resolution, i.e., measurements taken at one minute intervals. The volume of data gathered during a given period of investigation can be overwhelming, requiring significant time and effort to manually sift through it and exclude. Consequently, the automation of this process is considered to be highly advantageous.

One crucial element in the field of space weather physics pertains to the interrelationship among solar flare (SF) occurrences, the ionospheric reaction to such events, and Coronal Mass Ejections (CME) [2]. The importance of SF occurrences, including those classified as lower M-class rather than X-class, holds considerable significance and is currently a research focus. SF events classified as M-class exhibit a lower correlation with CMEs compared to X-class flares; however, a subset of M-class flares has been found to be associated with faster CMEs [5], which aligns with the concept known as Big Flare Syndrome [5,6]. Moreover, there exists a significant correlation between X-class SFs of great intensity and CMEs, which have been observed to induce disruptions to satellite and navigation systems [7].

The current research in the fields of SFs, CMEs, and the ionosphere is focused on the application of machine learning (ML) techniques. Classification methods have been employed in the classification of lighting waveforms [8], as well as in the classification of radar returns [9–13] and auroral image classification [14,15]. Similar to the pre-processing phase of VLF data analysis, manual radar return classification entails human intervention and is a labor-intensive procedure [9]. The resemblance between the two aforementioned processes underscores the justification for employing ML classification techniques in order to automatically eliminate and/or classify erroneous ionospheric VLF data.

In order to streamline the manual exclusion or labeling of inaccurate ionospheric VLF data, the application of ML classification techniques has been considered. The aim of this study was to employ pre-existing ionospheric VLF amplitude data (Worldwide Archive Of Low-Frequency Data And Observations—WALDO, https://waldo.world/, accessed on 24 March 2023) that have been previously utilized in research, as well as soft range X-ray irradiance data (Geostationary Operational Environmental Satellite—GOES, https://www.ncei.noaa.gov/, accessed on 24 March 2023), with the purpose of investigating the feasibility of automatically labeling erroneous data and the effects of increased solar flare activity on measured VLF data. The task involved the utilization of the Random Forest (RF) method for ML classification. The data used for this task consisted of a total of 19 transmitter–receiver (T-R) pairs which are situated in North America. The data spanned the time period between September 2011 and October 2011, during which, solar flare activity was observed. In addition, in the manuscript, we discuss how the presented research serves as a potential method for the automated labeling of data in various fields of space science. The used datasets, results, and a post-processing workflow can be found on Zenodo: https://zenodo.org/record/8220971, accessed on 7 August 2023 (Supplementary Materials).

2. Materials and Methods

The present study made use of data obtained in 2011, specifically data gathered during solar flare events that took place in September (ranging from C2.5 to X2.1) and October (ranging from C5.5 to M1.5). The data employed in this study had already undergone a process of "labeling", wherein erroneous data points were previously excluded (Figure 1). This circumstance presented a distinctive opportunity to utilize the dataset for the purpose of training a ML model, which, in turn, could potentially automate the process of labeling in subsequent instances. Figure 1 illustrates four instances in which a researcher was required to manually exclude erroneous data points from the dataset in preparation for a subsequent analysis. Figure 1a illustrates the impact of noisy data or errors in instrumentation on VLF data. Conversely, Figure 1b demonstrates the effects of SFs on VLF data, including the presence of outlier data points (single data points that significantly deviate from the rest of the measured data points). Figure 1c,d exhibit a confluence of SF impacts, anomalous data

points, and instrumentation errors. The aforementioned instances have a direct impact on the processing time required, i.e., the "data cleaning". This is particularly significant due to the high measurement rate and the time spans of VLF data. Consequently, this may result in large datasets and necessitate a substantial amount of time for the manual exclusion of such data.

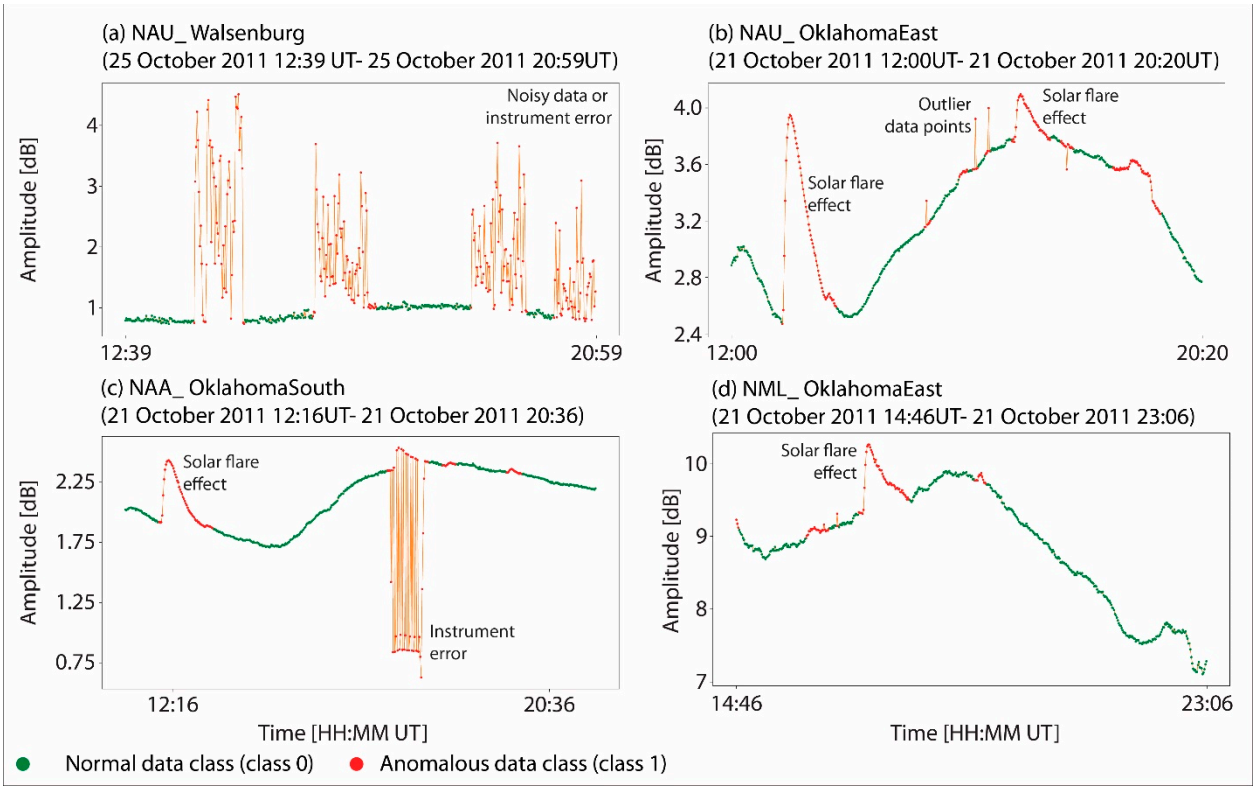

Figure 1. Illustrations of erroneous VLF data; (**a**) example of instrumentation error or noisy data; (**b**) example of solar flare effects on VLF data with outlier data points; (**c**) example of solar flare effects and instrumentation error; and (**d**) example of solar flare effects and outlier data points.

The dataset employed in this study comprised five VLF transmitters, namely NPM, NLK, NML, NAA, and NAU, along with four VLF receivers, specifically Walsenburg, Oklahoma South, Oklahoma East, and Sheridan (Figure 2). There was a total of 19 T-R pairs.

In standard ML workflows, a crucial step involves the pre-processing of data (Figure 3). During this phase, the data are transformed in order to meet to the requirements of the ML methods, i.e., dataset formats. Additionally, features are extracted, which will serve the purpose of revealing important patterns and characteristics within the dataset. Furthermore, the complete dataset is divided into separate training and testing datasets. In cases where the classes within the training dataset are imbalanced, it is common practice to balance them. This balancing process enhances the evaluation metrics and predictive capabilities of the model.

The samples, in which the excluded data points were appropriately labeled, were merged with the original dataset, comprising both the excluded and non-excluded data points. This integration resulted in the creation of a unified database, serving as the foundation for the ML modeling process. The database was updated with soft-range X-ray irradiance data [16], VLF transmitter, and VLF receiver information [17], as well as local receiver time data. The primary features of the database were the VLF amplitude data and X-ray data. These features served as the basis for calculating the other features. Additionally, the transmitter, receiver, and local receiver time were considered as secondary features and played crucial roles in establishing the core of the database. The target variable was encoded as binary data, where the data points from the labeled samples that were

excluded were assigned a value of 1 in the target variable, indicating anomalous data points. Conversely, the data points that were retained in the labeled sample were assigned a value of 0 in the target variable, representing normal data points. A process of filtering the X-ray and VLF data was conducted to eliminate any data points that lacked a measured X-ray variable or VLF amplitude variable.

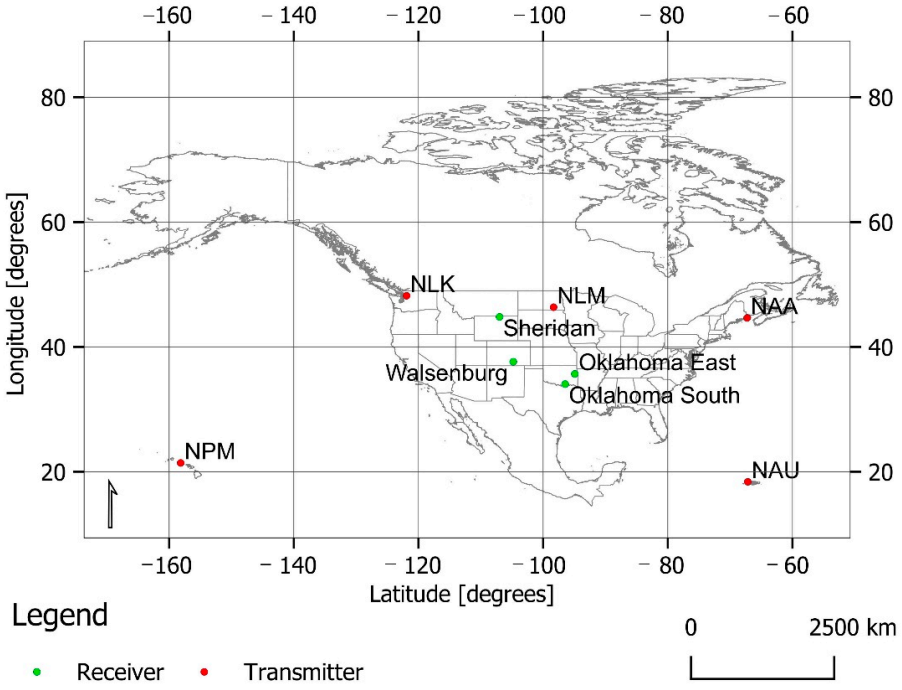

Figure 2. Spatial distribution of VLF transmitters and receivers utilized in this research (base map data obtained from DIVA-GIS, https://www.diva-gis.org/gdata, accessed on 1 April 2023).

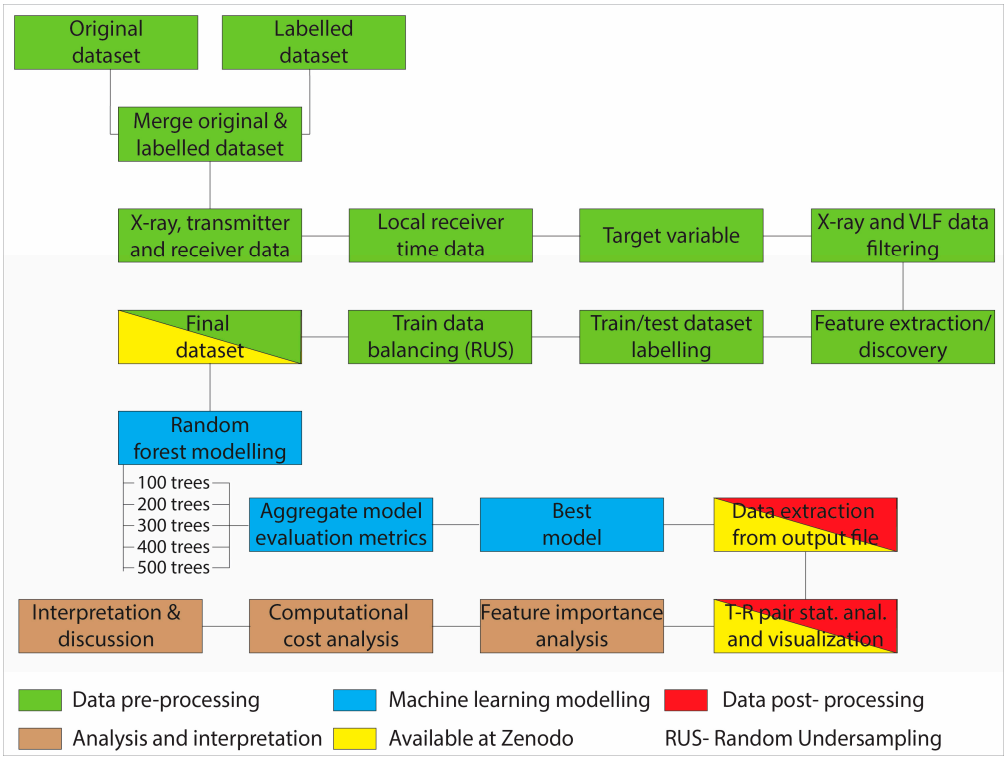

Figure 3. Pre-processing, modelling, and post-processing workflow.

The process of feature extraction, also referred to as discovery, pertains to the identification of features which are denoted as tertiary features in this study. The tertiary features were calculated based on the primary features, namely the VLF amplitude and X-ray data. The tertiary features were classified as statistical features, as they contained relevant information regarding the rolling window statistics. These statistics included the standard deviation, mean, and median values of the rolling window, which were computed for various window sizes. Specifically, the window sizes considered were 5 (short-term), 20 (mid-term), and 180 (long-term) minutes, representing different time dependencies within the dataset. Furthermore, the data were augmented by incorporating a lagged signal for time intervals ranging from 1 to 5 min. Additionally, the rate of change, as well as the first and second differentials, were calculated for the primary features' data. The most recent additions corresponded to a set of binary features, which encoded whether a given data point exceeded the mean or median value of the VLF amplitude data. The primary objective of the tertiary features was to determine whether any statistical parameters contained valuable information for the ML classification task. The total number of features was 41, as tertiary features were computed for both the VLF amplitude data and X-ray data, in addition to those exceeding the mean or median values. An overview of these features is presented in Table 1.

Table 1. Hierarchical feature classification for RF modelling.

Feature	Class	Class 2	Feature	Class	Class 2
VLF amplitude	P	Measured	RW median	T	Calculated
X-ray	P	Measured	Lagged signal	T	Calculated
Receiver local time	S	Measured	Rate of change	T	Calculated
Transmitter	S	/	First difference	T	Calculated
Receiver	S	/	Second difference	T	Calculated
RW st. dev.	T	Calculated	Higher than mean	T	Calculated
RW mean	T	Calculated	Higher than median	T	Calculated

RW—Rolling window; st. dev—Standard deviation; P—Primary; S—Secondary; and T—Tertiary.

Once the features were generated for the entire database, it was divided into separate training and testing databases. Subsequently, the training and testing data points were appropriately labeled to facilitate their utilization in the JASP software [18]. The last stage of data pre-processing involved addressing the imbalance in the training dataset, as datasets with imbalanced class distributions can introduce bias towards the majority class [19–21]. Imbalanced classification tasks pertain to an inherent disparity or disproportion between classes in binary classification problems. The methods employed to address this class imbalance can be categorized into two main approaches: under-sampling and oversampling. Under-sampling involves applying methods to the majority class to reduce the number of instances, while oversampling entails techniques applied to the minority class to increase the number of minority instances. The present study employed the random under-sampling technique [22], which involves the random removal of instances from the majority class [23,24]. The issue of imbalanced classification is commonly observed in anomaly detection scenarios, as highlighted by [25]. In the present study, a similar situation arose, where one class represented the normal category, while the other class pertained to the anomalous data category, specifically the excluded data class.

The Random Forest (RF) algorithm was proposed by Breiman in 2001 [26]. It has been extensively utilized in various scientific and industrial domains for classification and regression tasks over the past two decades. The RF algorithm is known for its ability to avoid overfitting due to the implementation of averaging or voting [27] and the utilization of the law of large numbers [26]. This algorithm has gained significant popularity due to its simplicity, as it only requires the specification of the number of trees (the decision trees used to control the complexity of the model) as a hyperparameter.

The initial RF classification in this research was conducted using five models, distinguished solely by the varying number of trees employed. The quantity of trees varied between 100 and 500, increasing by increments of 100 trees. The evaluation metrics of the aggregate model were analyzed to determine the optimal model. In cases where there was no clear best model, the model parsimony method could be employed. This method selected the model with the fewest hyperparameters, indicating that simplicity was a desirable quality in the best model.

The study employed several classification evaluation metrics, including accuracy, precision, recall, false positive rate, AUC, F1-score, Matthew's correlation coefficient (MCC), and statistical parity parameter.

(a) The accuracy parameter can be defined as the proportion of instances that were correctly classified out of the total number of instances, encompassing both true positive and true negative classifications.
(b) The precision parameter is defined as the proportion of correct positive predictions out of the total number of predicted positives, specifically, the ratio of true positives to the sum of true positives and false positives.
(c) The recall parameter, which is alternatively referred to as the true positive rate, quantified the proportion of correctly identified positive instances in relation to the combined count of true positives and true negatives. In other words, recall assessed the fraction of positive instances that were accurately classified [28].
(d) The false positive rate refers to the proportion of incorrect positive predictions relative to the overall number of instances in the negative class. Specifically, it quantified the ratio of misclassified instances that were predicted as being in the negative class, but were actually part of the positive class, to the total number of instances in the negative class.
(e) The Area Under the Receiver Operating Characteristic Curve (AUC) is a commonly employed, single-number metric for evaluating classification performance [29]. It is suggested as a comprehensive measure in ML classification tasks [30], and is a widely used measure for assessing the overall classification performance of a classifier [28]. Compared to accuracy, the AUC is considered to be a superior metric [30]. Moreover, the AUC parameter was employed to determine the model's ability to differentiate between the positive and negative classes, which refers to the model's discrimination performance [31]. From a practical standpoint, AUC values of 0.5 indicated that the model lacked the ability to distinguish between classes and performed random classification. Conversely, AUC values closer to 1 were more desirable, since they represent better classification models.
(f) The F1-score evaluation metric calculated the harmonic mean between the recall (true positive rate) and precision. In the context of imbalanced binary classification tasks, the F1-score has been identified as a more desirable metric compared to accuracy [28,32].
(g) The Matthew's correlation coefficient (MCC) is considered to be an even more favorable alternative to both F1-score and accuracy, due to its reliability and the requirement for a high MCC score to indicate a successful performance across all four categories of the confusion matrix (true positive, false positive, true negative, and false negative) [33]. Moreover, it has been observed that both accuracy and F1-score can exhibit inflated values compared to the actual value when dealing with imbalanced datasets [33], and that the MCC should be used as a standard for imbalanced datasets [34].
(h) The statistical parity parameter, which is the simplest parameter out of all those mentioned earlier, indicated the proportion of classified instances per-class out of all the classified instances. In other words, the per-class values of the statistical parity parameter should closely align with the actual distribution of the data in the test set for a good classification model. The metrics that were previously presented were employed based on specific requirements, and furthermore, on a per-class basis, to ensure a comprehensive statistical analysis.

3. Results

The workflow consisted of multiple stages in the processing of the data, with the primary stages being: the pre-processing of the data (including the division of the data into training and testing sets, as well as balancing the classes within the training dataset), modeling using RF with different numbers of trees, the selection of the best model, and the evaluation of the metric statistics for each T-R pair. These stages are presented in more detail in the following sections.

3.1. Data Pre-Processing

The research utilized a full set of training data consisting of 19 T-R pairs that collected data during solar flare events in September 2011, measured with 1 min intervals. The dataset comprised a total of 135,308 data points prior to balancing, with a class distribution of 22% for the anomalous data (designated as 1 in the database) and 78% for the normal data (class 0). Following the implementation of the random under-sampling technique, the class distribution achieved a balanced state, with an equal distribution of 50% for each class. The resulting training dataset consisted of a total of 59,344 data points. The database's testing phase encompassed solar flare events that were documented in October 2011. It consisted of 19 T-R pairs and a total of 180,071 data points.

3.2. Random Forest Modelling

The RF modeling was conducted using a range of trees, ranging from 100 to 500, with increments of 100 trees. Figure 4 depicts the accuracy, precision, F1-score, and AUC parameter for all five models. The values of each model for all four evaluation metrics exhibited a notable degree of similarity. The out-of-bag classification accuracy demonstrated convergence in the random forest model with 50 trees, suggesting that the subsequent models would yield comparable outcomes. The RF model, consisting of 100 trees, was selected as the optimal model for the study. The initial rationale behind this observation was that, despite the close resemblance of the four evaluation metrics, the RF model with 100 trees exhibited marginally superior outcomes in terms of accuracy and F1-score. Another factor considered was that models with a smaller number of trees require less computational resources, i.e., model parsimony. The ideal model is the one that utilizes the fewest (hyper)parameters and is the simplest. Furthermore, the statistical parity evaluation metric revealed that the RF model, employing a total of 100 trees and utilizing a balanced dataset, exhibited the greatest degree of success in predicting the class ratio within the testing dataset. Specifically, the RF model correctly identified 16.3% of the instances as anomalous, which closely aligned with the actual proportion of 15% anomalous instances present in the testing dataset. The other RF models exhibited a range of anomalous instances, with percentages varying from 16.7% to 17%.

In order to determine the optimal model, additional testing was conducted by varying the number of trees in the RF model. Specifically, the RF model was tested with 125, 150, and 175 trees, in addition to the initial model with 100 trees. The outcomes exhibited a high degree of comparability across the models, with minimal variations in their overall classification efficacy. Consequently, the RF model with 100 trees maintained its status as the superior model.

Within the testing dataset for the RF model, separation between various T-R pairs and the calculated evaluation metrics was conducted (Figure 5). The analysis involved interpreting the T-R pairs separately to enhance the understanding of the model's ability to distinguish between the normal and anomalous data points. The evaluation metrics of accuracy revealed that the T-R pair NML–Oklahoma South exhibited a distinct outlier status, with values of approximately 0.6. Regarding precision, both NPM–Walsenburg and NAA–Walsenburg were outliers, as they achieved a high precision score of approximately 0.94. In terms of the F1-score, NPM–Walsenburg exhibited relatively increased values. A median value of 0.87 for the F1-score suggested that half of the models outperformed a score of 0.87 when evaluated using the F1-score metric. An additional parameter employed

in this analysis was the MCC, which exhibited a varied range, spanning from 0.7 (NPM–Walsenburg) to 0.14 (NLK–Oklahoma South), with a median value of 0.45. Based on the analysis conducted using individual T-R pairs and focusing solely on the overall evaluation metrics, it can be concluded that NPM–Walsenburg exhibited the most favorable evaluation metrics among all the T-R pairs, whereas NML–Oklahoma South demonstrated the least favorable performance.

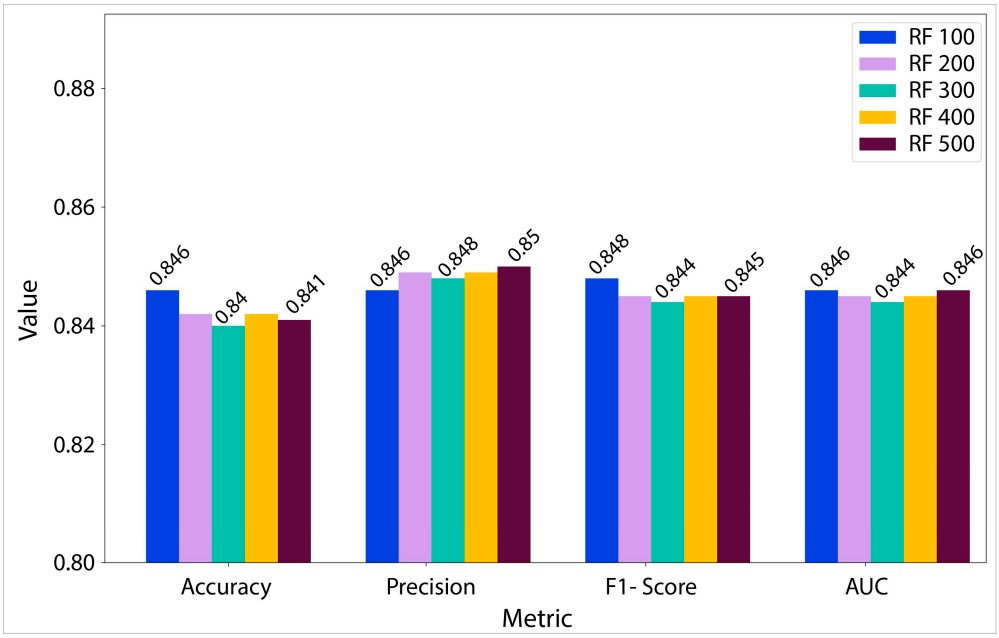

Figure 4. Comparison of Random Forest models with varying numbers of trees based on accuracy, precision, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

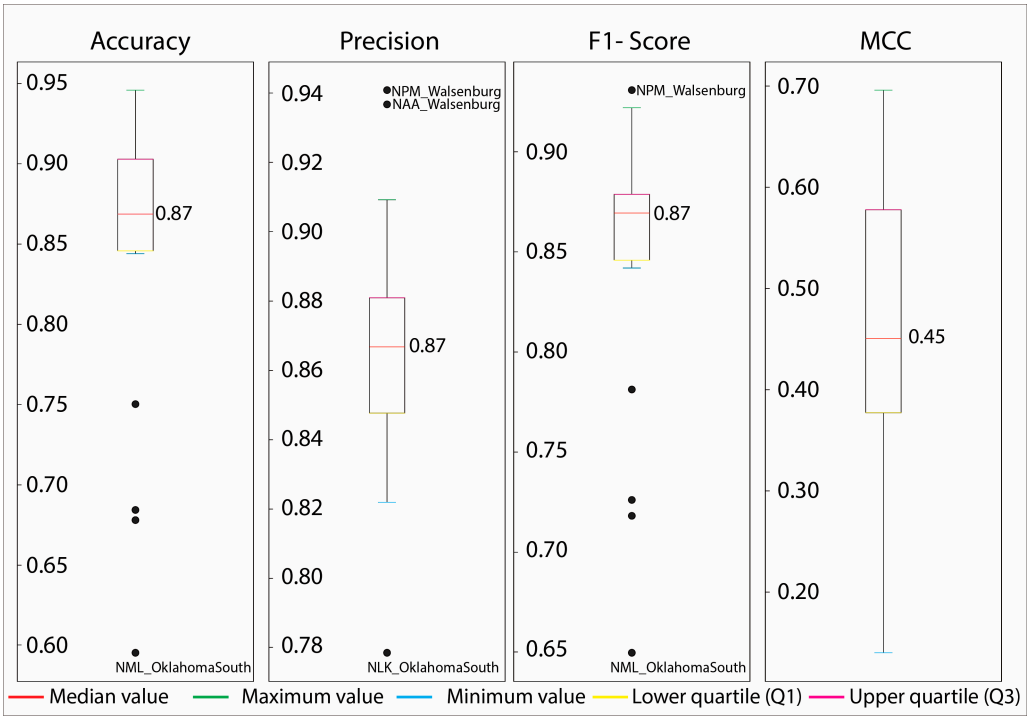

Figure 5. Accuracy, precision, F1-Score, and Matthew's correlation coefficient for individual transmitter–receiver pairs.

A more comprehensive examination was conducted by employing per-class evaluation metrics, which involved the calculation of the evaluation metrics for both the 0 (normal) and 1 (anomalous) classes within the testing database (Figure 6). The primary aim of computing the per-class evaluation metrics was to assess whether the model exhibited a significantly disparate capacity to predict one class compared to the others. Due to the imbalanced nature of this ML problem, the discrepancy may not be readily apparent when considering the overall evaluation metrics. Furthermore, the evaluation metrics for individual classes can offer additional understanding of the variability in the evaluation metrics and the effectiveness of the employed model in terms of classification.

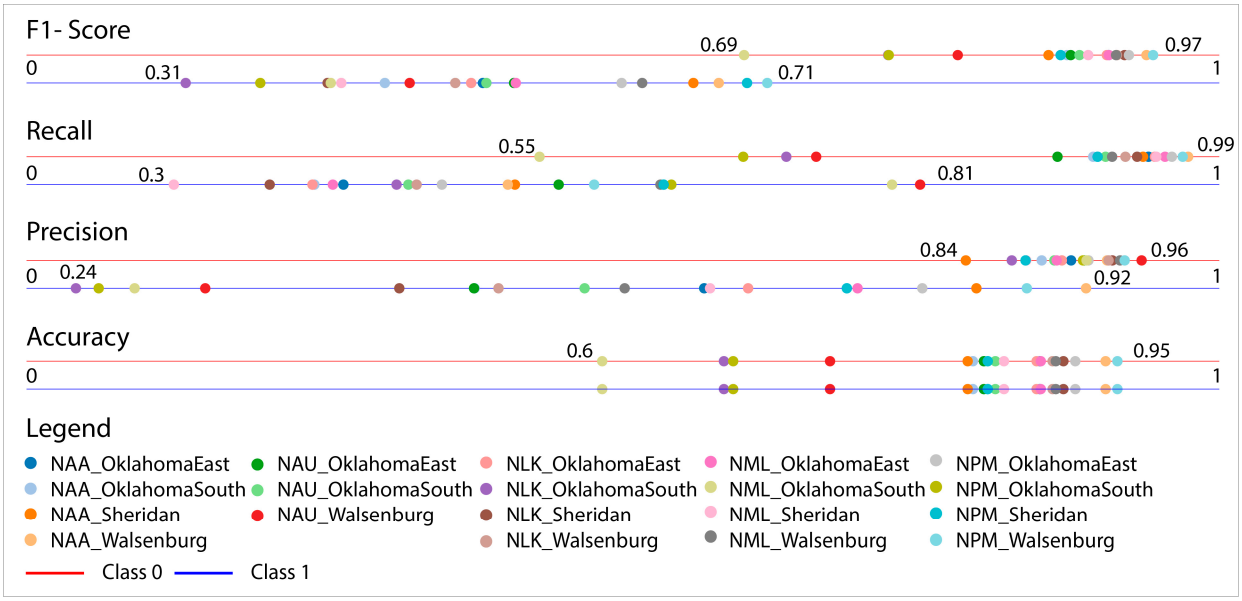

Figure 6. Per-class accuracy, precision, recall, and F1-Score values for individual transmitter–receiver pairs.

The F1-score offered initial insights by revealing that the range of values for class 0 spanned from 0.69 to 0.97, whereas the range for class 1 exhibited a lower minimum of 0.31 and a lower maximum of 0.71. This analysis offers initial observations regarding the classification efficacy of the model. Specifically, it revealed that the geometric mean of the true positive rate and precision was comparatively lower for the anomalous data class. For instance, the T-R pair NPM–Walsenburg achieved the highest overall F1-score of 0.94, as shown in Figure 5. However, the F1-score for the anomalous data class for the same T-R pair was lower at 0.71, whereas the F1-score for the normal data class was 0.97. Due to the imbalance in the ML task at hand, it was observed that the mean of both values was 0.94. Consequently, a comprehensive statistical analysis was warranted to further investigate this matter. In contrast, the F1-scores for class 1 and class 0 in the NML–Oklahoma South model, which is considered to be the poorest model according to Figure 5, were relatively low at 0.41 and 0.69, respectively. In all of the models that were constructed, it was observed that the F1-score for the anomalous data class was consistently lower than the F1-score for the normal data class. This discrepancy necessitated a per-class statistical analysis of the evaluation metric parameter due to the skewed distribution of the imbalanced problem in relation to the calculated evaluation metrics. The precision parameter revealed a notable disparity between the two classes. Specifically, class 0 exhibited a range of 0.12 (ranging from 0.84 to 0.96), whereas class 1 demonstrated a significantly broader range of 0.68. This discrepancy suggested that all the models achieved a relatively satisfactory precision range for the non-anomalous day class (class 0), but exhibited considerable variation in performance for the anomalous day class.

3.3. In-Depth Analysis of Selected Transmitter–Receiver Pair Classifications

Figure 7 gives an instance of the classification accuracy, showcasing the best two T-R pairs, namely NPM–Walsenburg and NAA–Walsenburg. In both visual representations, the upper panel presents X-ray irradiance data and the middle panel corresponds to the true labeling of the data, which was the manual classification performed by a researcher. Conversely, the lower panel represents the classification achieved through the utilization of an ML model.

The NPM–Walsenburg instance was selected due to its demonstration of an error in the signal measurement, which the model effectively detected and classified as anomalous data. It is important to acknowledge that not all the data points within the interrupted signal were categorized as anomalous; rather, a subset of data points was identified as normal. However, it is worth noting that only a limited number of instances of such occurrences were observed. Furthermore, during the start of the signal, a minor solar flare event was accurately identified as anomalous, aligning with the manual classification performed by the researchers. The classification of the amplitude signal as non-anomalous before the interruption of the signal and after the occurrence of the solar flare event was incorrect.

In contrast, the NAA–Walsenburg case study presented a collection of six solar flare events that were identified as anomalous. The RF model accurately detected five of these events, albeit with limited success in accurately determining their duration. Regrettably, one event was entirely omitted by the model.

Both instances exemplified real-world situations that researchers encounter during the processing of ionospheric VLF data, namely, signal interruption and the impact of solar flare events. In both scenarios, the RF model demonstrated a satisfactory classification performance. This approach allows researchers to save time by manually adjusting the data labels instead of fully processing the signal and labeling the data.

Figure 8a exhibits a time section commencing on 19 October 2011 at approximately 13:47 UT time and concluding on the same date at around 22:07 UT time. This time section encompassed a duration of 500 min, i.e., 500 data points. The results of the conducted manual labelling, which involved the interpretation of six anomalous time spans of the VLF amplitude signal, are presented in the middle panel of Figure 8a. Meanwhile, the RF classification of the signal resulted in a solitary, extensive anomalous region, which represented a wholly inaccurate classification of the signal. Consequently, the researcher would need to manually assign a label to the signal. The aforementioned observation can also be applied to the time interval depicted in Figure 8b, specifically that on the 20 October 2011. This interval exhibited a duration of 500 min, similar to the previous example. In this particular instance, a prominent solar flare event is observable in the middle panel of Figure 8b. The observed event was accurately recognized as anomalous, although it is worth noting that a significant portion of the signal was also classified as anomalous, indicating a discrepancy in the data labeling between the researcher and the RF model.

In both aforementioned instances, the inaccurate categorization performed by the RF model served as an illustration of the subpar classification that can also be observed from the model. In both of the instances, the utilization of automatic labelling was not feasible due to the model's inadequate classification performance, necessitating manual intervention.

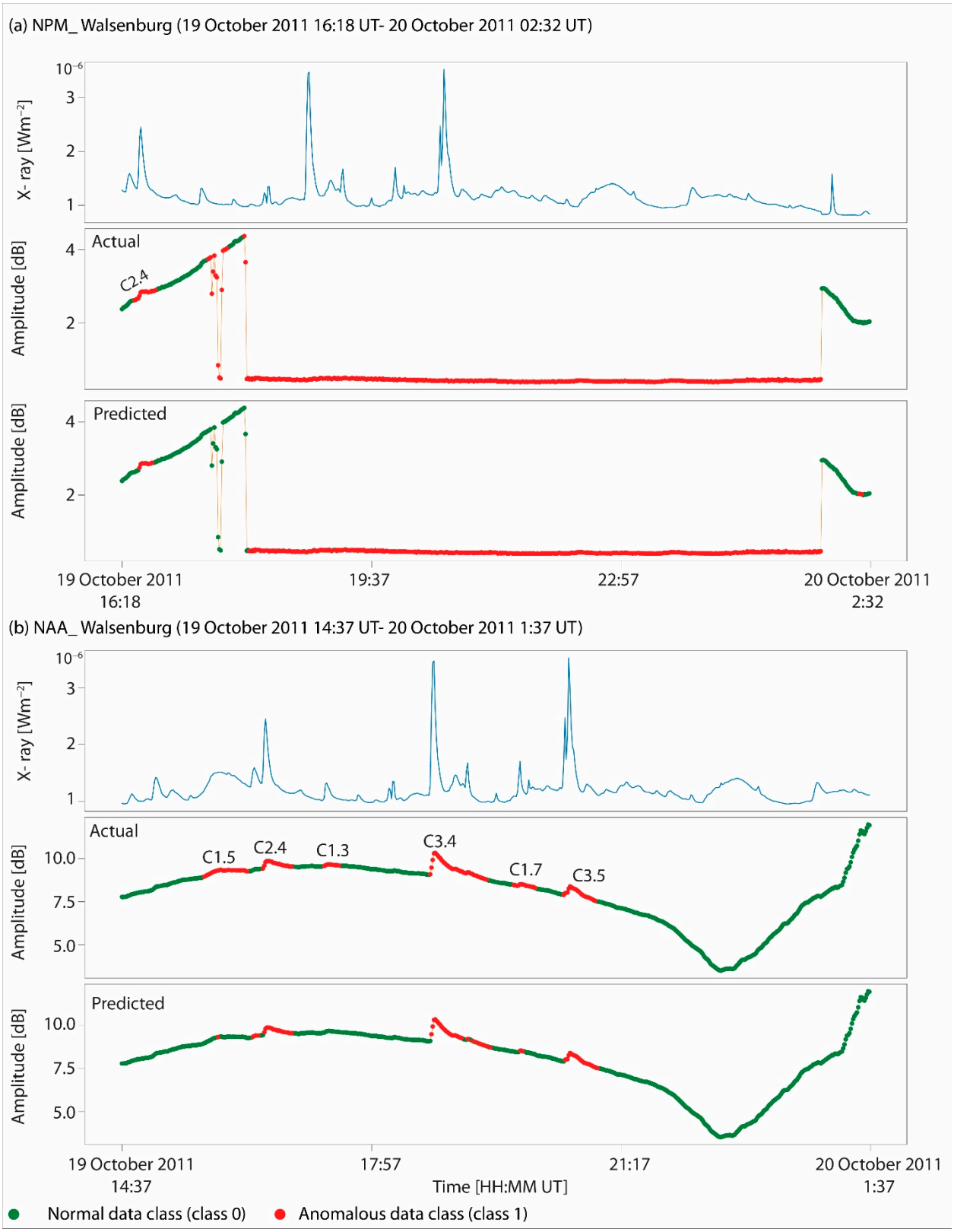

Figure 7. Examples of good classification; (**a**) transmitter–receiver pair NPM–Walsenburg, top panel—X-ray irradiance (from GOES), middle panel—true classification of the VLF amplitude signal, and bottom panel—output classification of the RF modelling; and (**b**) transmitter–receiver pair NAA–Walsenburg, top panel—X-ray irradiance, middle panel—true classification of the VLF amplitude signal, and bottom panel—output classification of the RF modelling.

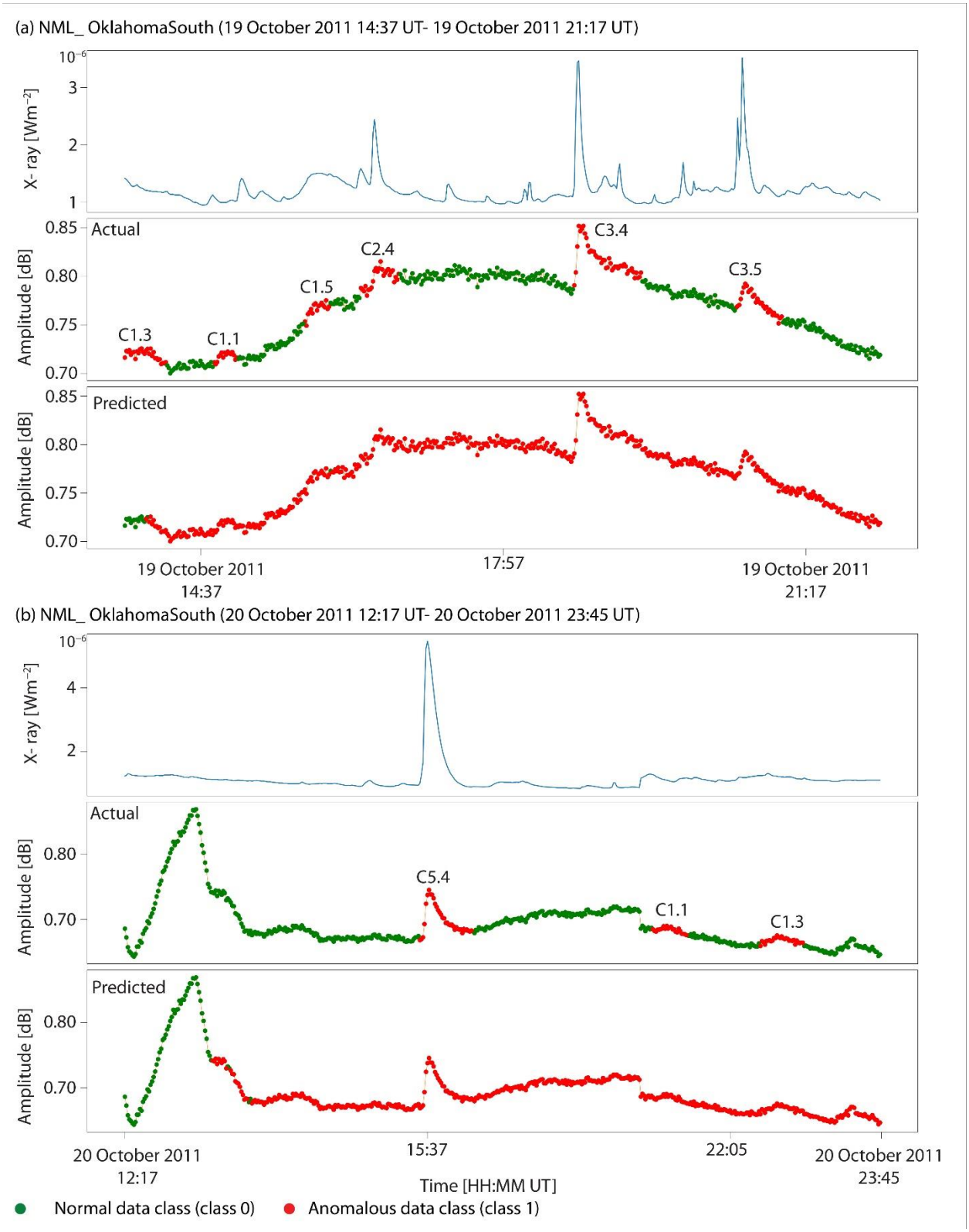

Figure 8. Examples of poor classification; (**a**) transmitter–receiver pair NML–Oklahoma South, top panel—X-ray irradiance (from GOES), middle panel—true classification of the VLF amplitude signal, and bottom panel—output classification of the RF modelling; and (**b**) transmitter–receiver pair NML–Oklahoma South, top panel—X-ray irradiance, middle panel—true classification of the VLF amplitude signal, and bottom panel—output classification of the RF modelling.

Model outputs, including solar flares of greater intensity, encompassing both M- and X-class events, are given in Figure 9. Figure 9a displays the T-R pair NAU–Oklahoma East on the 21 October from 11:59 to 20:19 UT. A total of six solar flares were observed, with five falling under the C class category, while one was classified as an M1.3 class solar flare. The RF model utilized in the analysis of the M1.3 solar flare demonstrated a reasonably accurate start time when the researcher opted to exclude the VLF data points. However, it exhibited a poorer classification of the end time for the excluded data. It is important to acknowledge that the M1.3 solar flare persisted and transitioned into a C1.6 solar flare, which was erroneously misclassified by the model. Regarding the additional solar flares depicted in Figure 9a, it is noteworthy that a C2.8 solar flare and a C.16 solar flare were both classified satisfactorily. Moreover, Figure 9a serves as a compelling illustration of outlier detection, showcasing four distinct occurrences of abrupt spikes (outlier data points) in the VLF signal. Notably, these instances were not accurately identified as anomalous by the RF model.

However, when the training and testing samples were rotated to enable the classification of X-class solar flares, the outcomes depicted in Figure 9b were achieved. Figure 9b illustrates an interesting instance and serves as a robust evaluation of the RF model. There were two primary factors contributing to the current event. Firstly, there was a significant solar flare of X-class (X2.1) magnitude. Secondly, there was a notable interruption in the signal, which persisted for a relatively extended period of time. The model successfully identified the majority of data points in the interrupted signal as anomalous. However, it incorrectly classified the beginning of the interrupted signal as non-anomalous. However, it should be noted that the model performed reasonably well in predicting the occurrence of an X-class solar flare. It accurately identified the beginning of the data exclusion period and correctly classified the entire signal as anomalous, with the exception of a slightly shorter duration of anomalous classification at the end compared to the researcher's classification.

Both of these examples provided much needed insights into the model's capabilities and limitations. Future research could potentially enhance this predictive power by conducting fine-tuning experiments and investigating various pre-processing techniques and other ML methods, which, in turn, could potentially provide better classification outcomes.

An examination of the feature importance was conducted subsequent to the determination of the extent of the capabilities exhibited by the best overall model. The findings from the analysis of the feature importance indicated that it was possible to develop a model with a reduced number of features (20) while still maintaining a relatively high level of predictive power and reducing the computational time (cost) by around 50%. For further details about the analysis of the feature importance, refer to Appendix A.

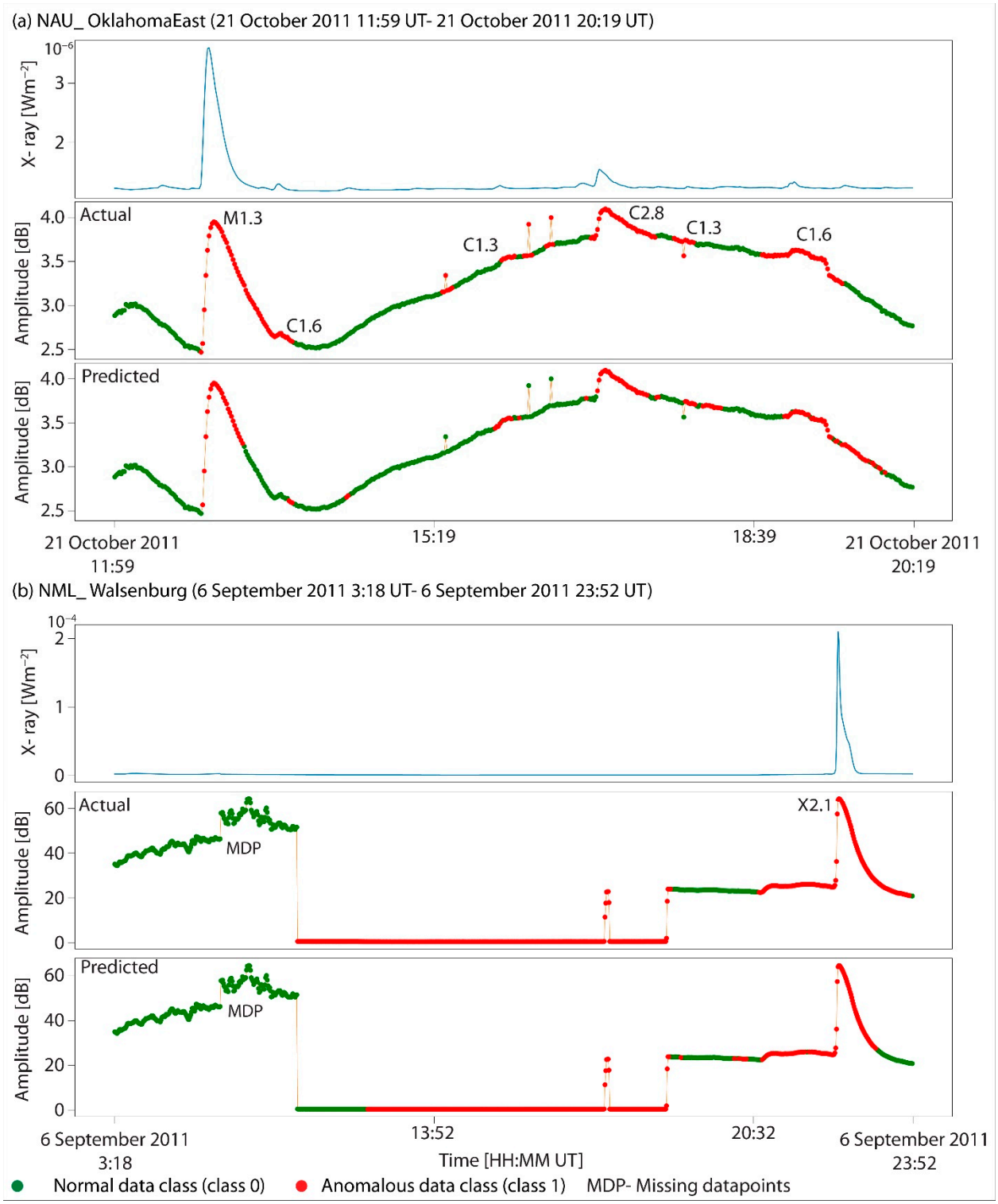

Figure 9. Examples of M- and X-class solar flares' classification; (**a**) transmitter–receiver pair NAU–Oklahoma East, top panel—X-ray irradiance (from GOES), middle panel—true classification of the VLF amplitude signal, and bottom panel—output classification of the RF modelling; (**b**) transmitter–receiver pair NML–Walsenburg, top panel—X-ray irradiance, middle panel—true classification of the VLF amplitude signal, and bottom panel—output classification of the RF modelling.

4. Discussion

The research presented in this study serves as an investigation into the potential for automating the manual labeling process of VLF ionospheric data. The statistical evaluation metrics produced models that have the potential for future refinement. However, for the purposes of this research paper, and when considering each individual T-R pair, there were several instances where the researcher would likely require minimal manual adjustments to the ML classification. Conversely, certain instances arose in which the RF classification produced entirely unsatisfactory outcomes, necessitating the researcher to undertake manual data relabeling. The utilization of ML techniques for the automated classification of VLF ionospheric data has the potential to reduce the amount of time researchers spend manually labeling and excluding data. However, it also presents opportunities for further research, which should aim to acquire supplementary data from a larger number of T-R pairs, as well as data from different time sections.

This endeavor has the potential to enhance the classification efficacy of the method. Furthermore, there are several other potential strategies that can be employed to enhance this classification accuracy. For instance, exploring alternative under-sampling and over-sampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) proposed by [35], could be beneficial. Unlike the random under-sampling method utilized in this study, SMOTE oversamples the minority class, thereby increasing the amount of data available in the training dataset. Furthermore, it is possible to evaluate various other ML classification techniques, including Support Vector Machines, K-nearest neighbors, and Artificial Neural Networks, among others. The aforementioned techniques have the potential to exert a substantial influence on the overall success rate of the method currently being investigated. Consequently, this could significantly reduce the amount of time researchers spend manually labeling and excluding erroneous data points from their datasets. Furthermore, the issue presented in this study pertained to a binary classification problem, wherein the two categories encompassed normal and anomalous data. Future research may also aim to address multi-class problems, specifically the classification of day and night VLF signals. This would involve a three-class problem, distinguishing regular and anomalous day and night signals. In conclusion, it is worth exploring the application of hybrid approaches that incorporate non-ML techniques, such as time series analyses and forecasting. These hybrid methods can be evaluated independently or in combination with ML methods, allowing for a direct comparison between ML and non-ML approaches, or the development of hybrid methodologies.

Regarding the RF method, it demonstrated its effectiveness and can be considered to be a highly favorable choice when addressing unfamiliar ML applications. In future research, it would be beneficial to conduct a comparative analysis of methods, even those deemed as "more complex" than the RF method.

5. Conclusions and Perspectives

Here, solar flares data, based on space- and ground-based observations, were investigated. As noted, the process of manually labeling solar flares data and subsequently excluding certain data is a demanding and laborious task that can be automated through the application of ML techniques. The present study employed the RF classification algorithm to effectively categorize anomalous data points, including instances of instrumentation errors and the effects of solar flares. The primary findings of this research can be succinctly summarized as follows:

- The results of the RF classification analysis indicated that the model with 100 trees performed the best overall. However, it is worth noting that models with higher numbers of trees also demonstrated satisfactory evaluation metric statistics. The reason for selecting the RF model with 100 trees was due to the preference for a simpler model with fewer hyperparameters, as this requires less computational time.
- The per-class statistics indicated that the F1-scores for all T-R pairs were higher for the non-anomalous data class, ranging from 0.69 to 0.97, meanwhile the F1-scores

for the anomalous data class were lower, ranging from 0.31 to 0.71. Furthermore, the evaluation metrics employed for the per-class analysis, specifically the recall and precision, exhibited similar characteristics, namely, lower values of the metrics for the anomalous data class. This suggested that the model exhibited a relatively weaker ability to accurately classify instances belonging to the anomalous data class compared to those belonging to the non-anomalous data class.

- Instances of effective classification were demonstrated through the NAA–Walsenburg and NPM–Walsenburg T-R pair examples. Both examples demonstrated satisfactory classifications, accompanied by satisfactory evaluation metric statistics. However, it should be noted that the T-R pair NML–Oklahoma South served as an illustrative case of poor classification. In this instance, the RF model incorrectly identified the majority of the time section as anomalous, thereby rendering the classification ineffective and requiring manual labeling.
- The presented research serves as an examination the potential of employing ML techniques for the automated labeling and/or exclusion of solar flares data.
- Future research should aim to investigate various oversampling and under-sampling techniques, as well as different ML methods, in order to determine if there is a more effective approach for this task. The RF algorithm yielded satisfactory outcomes in this context, affirming the feasibility of detecting VLF signal anomalies.
- As a perspective, this study could be used in the detection of the short-term responses of the ionosphere to gamma-ray bursts and in the analysis of pre-earthquake ionospheric anomalies, as well as in various fields of space science, where signals and data are very difficult to examine.
- Additional perspectives from this study can be found in broader applications of ML classification (or regression) methods to ionospheric VLF data. One potential application pertains to the automatic detection of daytime and nighttime signals. This involves the utilization of classification methods to address multi-class problems, which include the classification of normal daytime signals, anomalous daytime signals, nighttime signals, and transition zones related to terminators. Applying such methods and testing other use cases might yield significant insights into the issues pertaining to VLF ionospheric research and data.

Supplementary Materials: The training and testing datasets used in this study are available online at: https://zenodo.org/record/8220971, accessed on 7 August 2023.

Author Contributions: Conceptualization, F.A. and A.K.; writing—original draft preparation, F.A. and A.K.; writing—review and editing F.A., A.K. and V.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Institute of Physics Belgrade, University of Belgrade, through a grant by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia.

Data Availability Statement: Publicly available datasets were analyzed in this study. National Centers for Environmental Information (NCEI) Available online: https://www.ncei.noaa.gov/. accessed on 24 March 2023. Worldwide archive of low-frequency data and observations (WALDO) Available online: https://waldo.world/, accessed on 24 March 2023.

Acknowledgments: VLF data are provided by the WALDO database (https://waldo.world, accessed on 1 January 2023), operated jointly by the Georgia Institute of Technology and the University of Colorado Denver, using data collected from those institutions as well as Stanford University, and has been supported by various US government grants from the NSF, NASA, and the Department of Defense.

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

Appendix A. Feature Importance Analysis

Appendix A.1. Methodology

The final stage of the ML modeling phase involved conducting an analysis of the feature importance and evaluating the computational costs associated with the model. The analysis of the feature importance was conducted using the parameters of Mean Decrease in Node Accuracy (MDNA) and Total Increase in Node Purity (TINP). The MDNA parameter quantified the average reduction in the model's accuracy for each individual feature, while the TINP parameter quantified the average increase in the node purity for each individual feature in the model. In order for a feature to be deemed significant, it should maximize both the MDNA and the TINP. The ranking of the features and the analysis of the computational costs were conducted based on the scores of the MDNA and TINP. The number of features was systematically reduced from the full set of 41 to subsets of 20, 10, and 5. The duration of training and testing the model was recorded to assess the time taken. The examination of the evaluation metrics and the computational time required by the model was considered as a computational cost analysis, which involved balancing the trade-off between maximizing certain evaluation metrics and the time needed for the model training and testing. The reduction in the computational time, while maintaining similarity to the evaluation metrics of the model that used all features, was considered advantageous for future research.

Appendix A.2. Results

The analysis of the feature importance was a crucial step in refining the model, eliminating irrelevant features, and optimizing the overall performance of the model, which was derived from the RF modelling process. Here, the MDNA and TINP parameters were employed to determine the feature importance. Figure A1 presents a visual representation of the top 20 most informative features. Among the various features examined, it was found that both the MDNA and TINP yielded valuable insights. Notably, the local receiver time emerged as the most informative feature, followed by a statistical measure termed the rolling standard deviation of the VLF amplitude. This particular measure utilized the preceding 180 data points for its analysis. An additional noteworthy observation from the feature importance ranking was that the rolling window statistics comprised 60% (12 out of 20) of the top 20 most informative features. Figure A1 did not exhibit certain features, which were regarded as the least informative. The list of features with low information content included the second derivative of the X-ray signal, the binary class that exceeded the mean or median, and the first and second derivatives of the VLF amplitude and X-ray data.

The sequential RF modeling approach was employed based on the insights obtained from the feature importance analysis. This involved constructing a series of models using the same methodology as that of the previous best model, which utilized 100 trees in the RF model. However, the subsequent models were constructed with reduced numbers of features. Specifically, the models were created using the 20 most informative features (half of the original set), the 10 most informative features (a quarter of the original set), and the 5 most informative features (an eighth of the original set). The model was subsequently compared to the previous best model, which was the RF model that utilized all 41 features. Additionally, a computational cost analysis was conducted to evaluate the model's training and testing time.

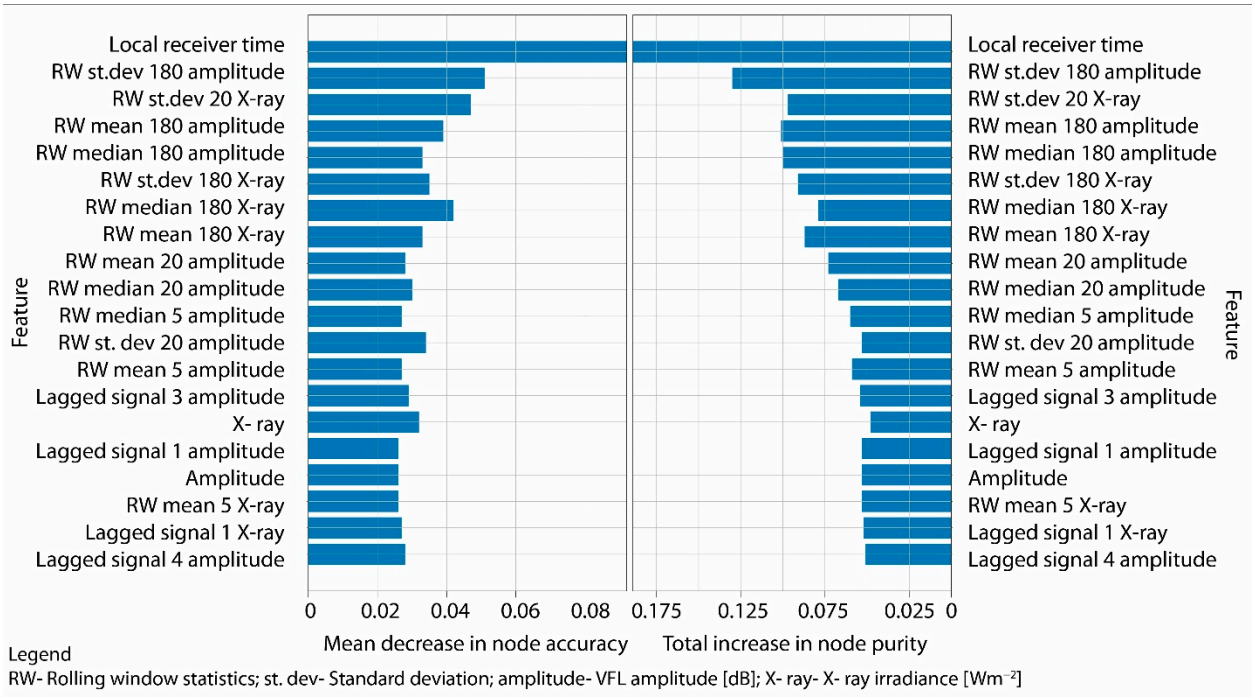

Figure A1. Top 20 most informative features from the RF modelling based on the mean decrease in node accuracy (**left**) and total increase in node purity (**right**).

Table A1 presents the model's performance using different feature sets, including the full set of 41 features, as well as subsets of 20, 10, and 5 features. The evaluation metrics considered for each class included the F1-score, AUC, recall, false positive rate, precision, accuracy, statistical parity, and the duration of training and testing the model. In terms of the F1-score, the model that utilized all 41 features demonstrated the highest performance for the anomalous data class, achieving a value of 0.508. This observation also held true for the AUC parameter, which showed the performance of both normal and anomalous data scores. The recall values, particularly for the anomalous data display, indicated that the model employing the fewest features exhibited an unexpectedly higher recall value compared to the other models, with a value of 0.663. The relationship between the precision and accuracy metrics was not consistent in the model that used the fewest features. Specifically, the evaluation metrics for the anomalous data class showed a decreasing trend as the number of features decreased. The statistical parity parameter revealed that the model with the lowest number of features exhibited an overestimation of the anomalous data class by approximately 9.7%, representing the most significant disparity observed among all the constructed models. In general, the model that employed the least number of features may be rejected, despite exhibiting the highest recall value. This was because it significantly overestimated the anomalous data class to a greater extent compared to the other models. In light of the assessment criteria, a choice can be reached regarding the models that employed 20 and 10 features. Both models exhibited the highest difference in the F1-scores for the anomalous data class, with a value of 3.3%. In the case of the three models, the difference between the highest and lowest AUC values for both classes were only slightly above 1%. Additionally, the highest difference in the false positive rate for the anomalous data class was recorded at 1.9% between all three models. The primary distinction was observed in the training duration of the models. Specifically, the model employing the complete feature set required slightly over two minutes (124 s) to complete the training and testing processes. In contrast, the model utilizing 20 features necessitated approximately half the time (59 s). Similarly, the model employing only 10 features required approximately 34 s for the training and testing phases.

Table A1. Selected evaluation metric statistics for models utilizing different numbers of features and the Random Forest model with 100 trees.

EM/No. Features	Class	41 Features	20 Features	10 Features	5 Features
F1-score	0	0.908	0.904	0.899	0.876
	1	0.508	0.475	0.489	0.502
AUC	0	0.847	0.839	0.844	0.84
	1	0.845	0.836	0.841	0.838
Recall	0	0.902	0.899	0.883	0.827
	1	0.53	0.489	0.538	0.663
FP Rate	0	0.47	0.511	0.462	0.337
	1	0.098	0.101	0.117	0.173
Precision	0	0.915	0.909	0.915	0.933
	1	0.488	0.461	0.449	0.404
Accuracy	0	0.846	0.837	0.831	0.802
	1	0.846	0.837	0.831	0.802
Statistical Parity *	0	0.837	0.84	0.82	0.753
	1	0.163	0.16	0.18	0.247
Time	/	124	59	34	24

* The true class distributions in the test dataset were 0.85 for class 0 and 0.15 for class 1; EM—evaluation metric.

As a result, the most optimal solution was achieved by employing models that incorporated only half or a quarter of the features present in the complete set of models. The evaluation metrics presented were highly comparable, with minimal compromises made in terms of statistical integrity, while significantly reducing the time required for the training and testing of the model (by 50% or 72%). In this particular instance, the disparity did not result in a significant time reduction in absolute terms (from approximately 65 to 90 s). However, when working with a substantially larger volume of data, this discrepancy could lead to a more substantial time reduction, consequently reducing the computational expenses.

References

1. Barta, V.; Natras, R.; Srećković, V.; Koronczay, D.; Schmidt, M.; Šulic, D. Multi-instrumental investigation of the solar flares impact on the ionosphere on 05–06 December 2006. *Front. Environ. Sci.* **2022**, *10*, 904335. [CrossRef]
2. Kolarski, A.; Veselinović, N.; Srećković, V.A.; Mijić, Z.; Savić, M.; Dragić, A. Impacts of Extreme Space Weather Events on September 6th, 2017 on Ionosphere and Primary Cosmic Rays. *Remote Sens.* **2023**, *15*, 1403. [CrossRef]
3. Grubor, D.P.; Šulić, D.M.; Žigman, V. Classification of X-ray Solar Flares Regarding Their Effects on the Lower Ionosphere Electron Density Profile. *Ann. Geophys.* **2008**, *26*, 1731–1740. [CrossRef]
4. Kolarski, A.; Srećković, V.A.; Mijić, Z.R. Response of the Earth's Lower Ionosphere to Solar Flares and Lightning-Induced Electron Precipitation Events by Analysis of VLF Signals: Similarities and Differences. *Appl. Sci.* **2022**, *12*, 582. [CrossRef]
5. Miteva, R.; Samwel, S.W. M-Class Solar Flares in Solar Cycles 23 and 24: Properties and Space Weather Relevance. *Universe* **2022**, *8*, 39. [CrossRef]
6. Kahler, S.W. The Role of the Big Flare Syndrome in Correlations of Solar Energetic Proton Fluxes and Associated Microwave Burst Parameters. *J. Geophys. Res.* **1982**, *87*, 3439. [CrossRef]
7. Srećković, V.A.; Šulić, D.M.; Vujčić, V.; Mijić, Z.R.; Ignjatović, L.M. Novel Modelling Approach for Obtaining the Parameters of Low Ionosphere under Extreme Radiation in X-Spectral Range. *Appl. Sci.* **2021**, *11*, 11574. [CrossRef]
8. Wang, J.; Huang, Q.; Ma, Q.; Chang, S.; He, J.; Wang, H.; Zhou, X.; Xiao, F.; Gao, C. Classification of VLF/LF Lightning Signals Using Sensors and Deep Learning Methods. *Sensors* **2020**, *20*, 1030. [CrossRef]
9. Sigillito, V.; Wing, S.; Hutton, L.; Baker, K. Classification of Radar Returns from the Ionosphere Using Neural Networks. *Johns Hopkins APL Tech. Dig.* **1989**, *10*, 262–266.
10. Dhande, J.; Dandekar, D.R. PSO Based SVM as an Optimal Classifier for Classification of Radar Returns from Ionosphere. *Int. J. Emerg. Technol.* **2011**, *2*, 1–3.
11. Oo, A.N. Classification of Radar Returns from Ionosphere Using NB-Tree and CFS. *Int. J. Trend Sci. Res. Dev.* **2018**, *2*, 1640–1642. [CrossRef]
12. Ameer Basha, G.; Lakshmana Gupta, K.; Ramakrishna, K. Expectation of Radar Returns from Ionosphere Using Decision Tree Technique. In *Advances in Data Science and Management*; Springer Nature: Singapore, 2020; pp. 209–214. [CrossRef]
13. Adhikari, S.; Thapa, S.; Shah, B.K. Oversampling Based Classifiers for Categorization of Radar Returns from the Ionosphere. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020. [CrossRef]

14. Shang, Z.; Yao, Z.; Liu, J.; Xu, L.; Xu, Y.; Zhang, B.; Guo, R.; Wei, Y. Automated Classification of Auroral Images with Deep Neural Networks. *Universe* **2023**, *9*, 96. [CrossRef]
15. Lian, J.; Liu, T.; Zhou, Y. Aurora Classification in All-Sky Images via CNN–Transformer. *Universe* **2023**, *9*, 230. [CrossRef]
16. National Centers for Environmental Information (NCEI). Available online: https://www.ncei.noaa.gov/ (accessed on 24 March 2023).
17. Worldwide Archive of Low-Frequency Data and Observations (WALDO). Available online: https://waldo.world/ (accessed on 24 March 2023).
18. JASP—A Fresh Way to Do Statistics. Available online: https://jasp-stats.org/ (accessed on 1 April 2023).
19. Prusa, J.; Khoshgoftaar, T.M.; Dittman, D.J.; Napolitano, A. Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 13–15 August 2015. [CrossRef]
20. Kulkarni, A.; Chong, D.; Batarseh, F.A. Foundations of Data Imbalance and Solutions for a Data Democracy. In *Data Democracy*; Academic Press: Cambridge, MA, USA, 2020; pp. 83–106. [CrossRef]
21. Devi, D.; Biswas, S.K.; Purkayastha, B. A Review on Solution to Class Imbalance Problem: Undersampling Approaches. In Proceedings of the 2020 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2–4 July 2020. [CrossRef]
22. 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]
23. Hasanin, T.; Khoshgoftaar, T. The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. In Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 6–9 July 2018. [CrossRef]
24. Saripuddin, M.; Suliman, A.; Syarmila Sameon, S.; Jorgensen, B.N. Random Undersampling on Imbalance Time Series Data for Anomaly Detection. In Proceedings of the 2021 the 4th International Conference on Machine Learning and Machine Intelligence, Virtual, 1–3 December 2021. [CrossRef]
25. Mishra, S. Handling Imbalanced Data: SMOTE vs. Random Undersampling. *Int. Res. J. Eng. Technol.* **2017**, *4*, 317–320.
26. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random Forests for Classification in Ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef]
28. Hossin, M.; Sulaimani, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–11. [CrossRef]
29. Hand, D.; Till, R. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **2001**, *45*, 171–186. [CrossRef]
30. Jin, H.; Ling, C.X. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299–310. [CrossRef]
31. Rosset, S. Model Selection via the AUC. In Proceedings of the Twenty-First International Conference on Machine Learning—ICML, Banff, AB, Canada, 4–8 July 2004. [CrossRef]
32. Joshi, M.V. On Evaluating Performance of Classifiers for Rare Classes. In Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 9–12 December 2002. [CrossRef]
33. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
34. Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation. *BioData Min.* **2021**, *14*, 13. [CrossRef]
35. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]