Contrib. Astron. Obs. Skalnaté Pleso **55/2**, 13 – 23, (2025)
https://doi.org/10.31577/caosp.2025.55.2.13

Standardization framework of ionospheric Very Low Frequency (VLF) signal amplitude classes for machine learning-based anomaly detection: from calm ionospheric conditions to solar activity-induced dynamics

F. Arnaut, A. Kolarski, V.A. Srećković, M. Langović and S. Jevremović

Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia (E-mail: filip.arnaut@ipb.ac.rs)

Received: November 16, 2024; Accepted: November 25, 2024

Abstract. Machine learning (ML) techniques are extensively employed in the domain of near-Earth physics. An application of ML techniques is the anomaly detection of Very Low Frequency (VLF) ionospheric amplitude data. Prior research focused on the binary classification task, yielding promising results, and the subsequent exploration involves the multi-label classification of a broader spectrum of VLF amplitude signal features. This research paper introduces a standardization framework for labeling multi-class VLF amplitude features, including normal (daytime) signals, solar flare effects, nighttime signals, instrumental errors, and outlier data points. The primary aim of this standardization framework is to define all main VLF amplitude features, specify the conditions under which each VLF amplitude feature can be classified, and outline future initiatives for the development of additional tools to facilitate the labeling process. Future research will focus on developing supplementary tools and software packages for this purpose, with the ultimate objective of establishing a streamlined process from the Worldwide Archive of Low-Frequency Data and Observations (WALDO) database to labeled data and subsequently to ML models.

Key words: Ionosphere – D-region – solar flares – near-Earth physics – space weather

1. Introduction

The classification, notably the detection of various ionospheric Very Low Frequency (VLF) 3-30 kHz signal features, has been the subject of prior research (Arnaut et al., 2023, 2024a,b). The primary aim of the aforementioned research was to establish a data-driven methodology for the automatic detection of various ionospheric VLF signal variations. For those objectives, machine learning

(ML) techniques were the most appropriate. The utilization of ML techniques in near-Earth physics, including the classification of radar returns (Dhande & Dandekar, 2011; Oo, 2018; Ameer Basha et al., 2020; Adhikari et al., 2020), lightning signals (Wang et al., 2020), and auroral images (Shang et al., 2023; Lian et al., 2023), is extensively demonstrated. The prior work on anomaly detection on VLF amplitude signals pertained to a binary ML problem involving two classes: the normal data class, representing the undisturbed VLF signal, and the anomalous data class, encompassing various signal disturbances such as noise, instrumental errors, solar flare effects, nighttime signals, and others.

As the continuation of previously conducted research, the development of a multi-label methodology was proposed wherein each of the previously mentioned VLF signal features would operate independently of one group, unlike the binary classification conducted earlier. This methodology presents several advantages and disadvantages; notably, supervised ML techniques rely on a pre-classified training dataset, necessitating that a researcher perform the classification for model training. Conversely, the advantage lies in the potential creation of a fully automated method for detecting various VLF signal features, which could be developed into a near real-time system for monitoring VLF signal changes.

The quality of the training dataset is crucial for the successful development of supervised ML models, as these methods are data-driven. The largest repository of freely available narrowband VLF datasets is the Worldwide Archive of Low-Frequency Data and Observations (WALDO), which includes open datasets for specific time intervals and VLF transmitter-receiver pairs from 2005 to 2017. This extensive repository suffices for the successful development of the previously presented methodology (Arnaut et al., 2023, 2024a,b); however, for the initial attempt at transitioning from a binary to a multi-class ML problem, the previously employed dataset is adequate.

The primary aim of this research paper is to provide an overview and standardization of various VLF signal features, as data quality is vital; accurate data labeling is the initial step in ensuring satisfactory data quality. The research paper will propose various classifications of ionospheric VLF signal features, potentially covering main VLF signal variations. It will also aim to provide clear definitions for each of the classed of VLF signals, grounded in prior experiences with the binary ML methodology for VLF signal anomaly detection, with the objective of enabling a broader audience to label VLF amplitudes in a standardized way.

The standardization framework and future studies on the topic are important for the interdisciplinary domains of ML and ionospheric physics, and they also aid other researchers in understanding the Sun-Earth connection. The established methodology for automatic anomaly detection of VLF amplitude signals may be advantageous to researchers in various interdisciplinary fields related to the effects of solar flares on the environment, human health, and other domains.

2. Framework for VLF signal class standardization

Figure 1 presents the framework for classifying VLF signal changes. The proposed standardization framework delineates five distinct categories: normal (daytime) signal, nighttime signal, solar flare effects, instrumental error, and outlier data point(s). A comprehensive discussion and examples of each case will be provided for each of the main five classes in the subsequent sections of the paper.

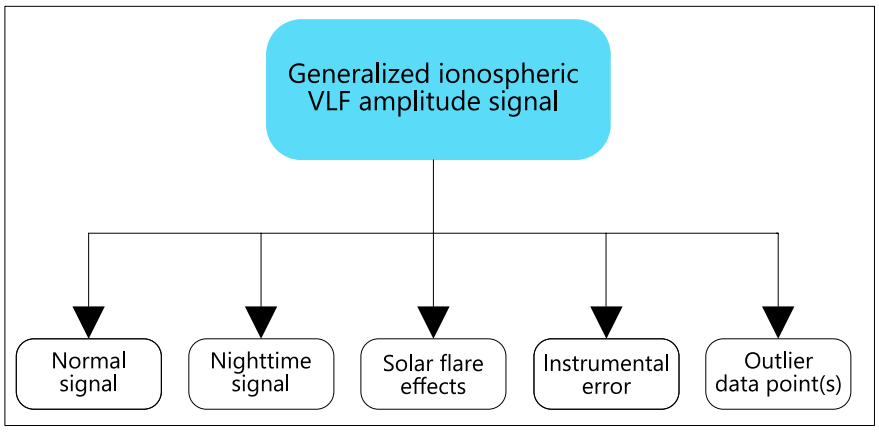

Figure 1. The proposed ionospheric VLF signal multi-label classification framework.

2.1. Normal daytime ionospheric VLF signal

The first and consequently simplest labeling category is the normal i.e., daytime ionospheric VLF signal, which consists of the undisturbed daytime signal. This signal is unaffected by extraterrestrial phenomena, such as solar flares, and is free from instrumental errors or anomalous data points. Figure 2 illustrates a signal in which the daytime readings are largely unaffected by extraterrestrial influences or the previously mentioned factors. The signal exhibits stability after the transitions from nighttime to daytime signal, which corresponds to sunrise, and from daytime to nighttime signal, which corresponds to sunset, stabilizing around 12 o'clock and fluctuating at its characteristic strength, which in presented case is around approximately 25 dB.

Figure 2 depicts an instance in which, despite significant fluctuations in the X-ray parameter, the variation in VLF amplitude is negligible. A decision must be made regarding the classification of these data points as belonging to the solar flare category. In the context of ML classification, such labeling may re-

sult in complications during modeling, as the model could erroneously classify undisturbed VLF amplitude values as solar flares, despite the absence of amplitude disturbances. The relatively undisturbed VLF amplitude signal, despite significant X-ray fluctuations, is not categorized under other classifications but is instead designated as part of the normal data class.

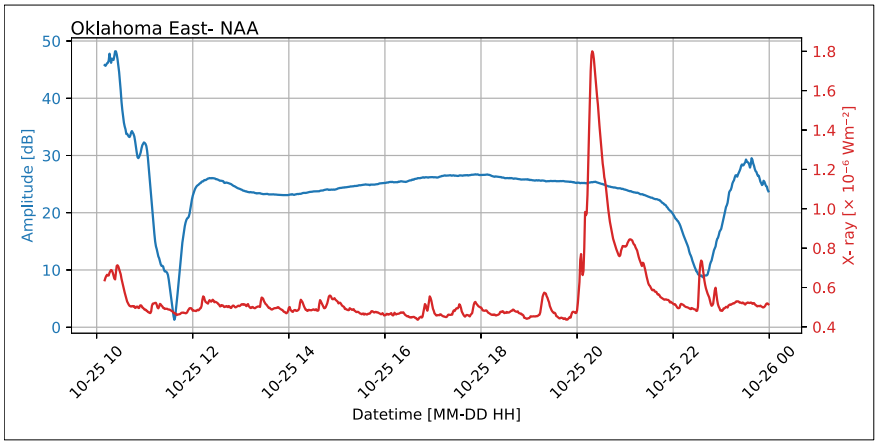

Figure 2. An example of a relatively undisturbed daytime signal on the Oklahoma East- NAA VLF amplitude signal; Blue line- VLF amplitude; Red line- X-ray irradiance data.

2.2. Nighttime signal

The nighttime signal in VLF amplitude appears as significantly elevated values in the VLF amplitude (Figure 3). When datasets correspond to multiple dates, a periodic component of the nighttime signal can be readily visually distinguished due to the characteristic that the nighttime signal exhibits greater values than the daytime signal. The nighttime signal is one of the more easily distinguishable VLF amplitude signal characteristics especially when there are more nighttime signals in a single data file.

The topic of discussion pertains to the appropriate initiation and completion points for labeling the nighttime VLF amplitude signal. Figure 3 illustrates two transitional periods: the transition from daytime to nighttime and the transition from nighttime to daytime (terminators). The terminators exhibit a unique characteristic wherein the transition from daytime to nighttime results in a signal reduction below daytime values, subsequently increasing to exceed daytime levels and stabilizing at a specific threshold. The transition from night to day is inverse, as the signal from the nighttime level decreases below the daytime

level before subsequently rising to the daytime level. The peak-to-peak value appears in both instances, and the labeling could split the peak-to-peak values for both transitional periods. Consequently, during the transition from day to night, the initial reduction of the peak-to-peak value would result in one half being classified as a daytime signal and the other half as a nighttime signal. The inverse is also applicable to the transition from nighttime to daytime.

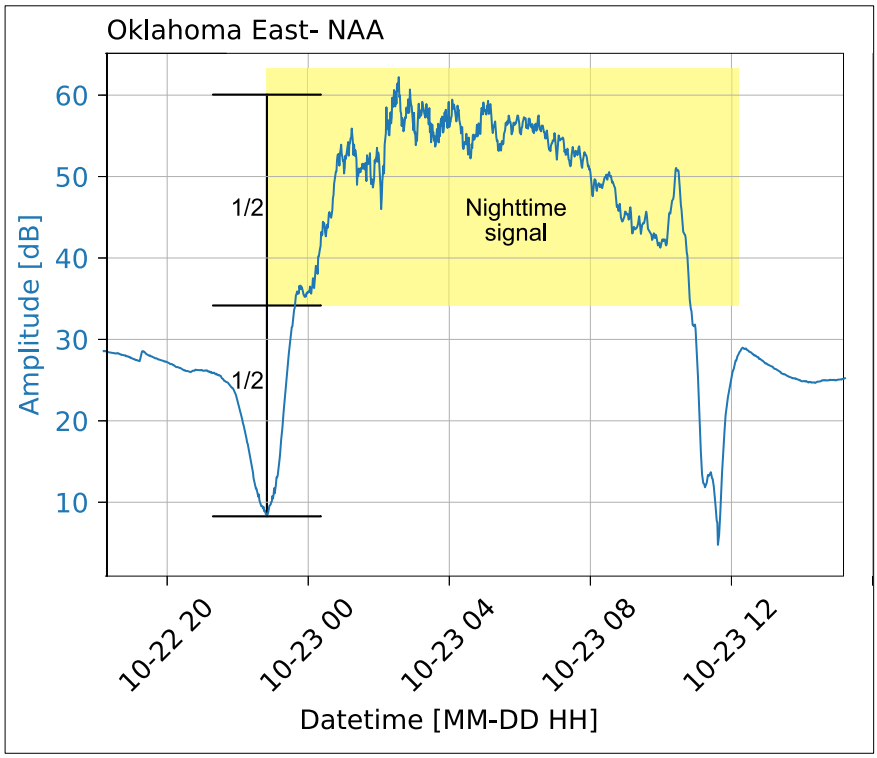

Figure 3. Example of the nighttime signal (yellow rectangle) and the transition of daytime-to-nighttime (marked by 1/2) and nighttime-to-daytime ionospheric VLF amplitude signal from the Oklahoma East- NAA transmitter- receiver pair; Blue line- VLF amplitude.

2.3. Solar flare effects

Solar flares are unique characteristics observable in the VLF amplitude signal. The solar flare effect is typically identifiable visually, characterized by a pronounced increase in the VLF amplitude signal, followed by a relatively gradual

return to pre-flare levels. Furthermore, plotting the X-ray irradiance data alongside the VLF amplitude facilitates a quite straightforward classification of solar flare effects on the VLF amplitude signal.

Figure 4 illustrates the Oklahoma East- NAA daytime signal featuring three distinct solar flare effects (purple rectangles) and a segment of the signal where the X-ray irradiance data suggests a solar flare; however, the VLF signal does not align with the expected increase in VLF amplitude (green rectangle). Labeling the pronounced solar flare effect is relatively straightforward when visual identification is combined with X-ray irradiance data; however, two questions emerge regarding the labeling process: when to initiate the labeling as solar flare effect data class and when to conclude it. For the standardization framework the solar flare class can initiate at the first data point that exhibits notable deviations from the preceding data point, where the overall morphology of the subsequent data points suggests a solar flare, in conjunction with the X-ray irradiance data. The conclusion of the solar flare classification can be determined when the data points revert to approximately pre-flare levels. Furthermore, it is important to acknowledge that the effects of solar flares are indistinguishable in nighttime VLF amplitude signals; thus, solar flare effects should only be identified in daytime signals.

Finally, the green rectangle illustrates a scenario in which the X-ray irradiance data exhibits a significant increase in values, whereas the VLF amplitude data remains unchanged. In the context of multi-label ML modeling, this should be considered a standard daytime signal, as no alterations are indicated in the VLF amplitude data despite a significant rise in the X-ray irradiance data. The primary aim of this standardization is to demonstrate to the model that not all increases in X-ray irradiance data correspond to an increase in VLF amplitude, indicating that the relationship between the two is not a one-to-one correlation. A similar situation can be seen on Figure 2 where the largest increase of X-ray irradiance does not result in an meaningful increase of VLF amplitude.

2.4. Instrumental errors

Instrumental errors in the VLF amplitude signal can manifest in various forms, but they are typically readily visually discernible. Figure 5a illustrates a scenario in which the VLF receiver fails to measure the VLF amplitude as either the transmitter or receiver was not functioning, resulting in the omission of observations. Typically, missing observations in the data file appear as observations lacking a numerical value; however, in Figure 5a, those observations are represented with a value of 0 (or any other constant value that is usually significantly lower than the real measured data values). If there is no measured real value, then the zero values would be absent, and only the signals preceding and succeeding them would be observable with a gap in the middle.

Figure 5b illustrates a scenario in which the signal exhibits erroneous VLF amplitude values, characterized by patterns of rapid increases and decreases in

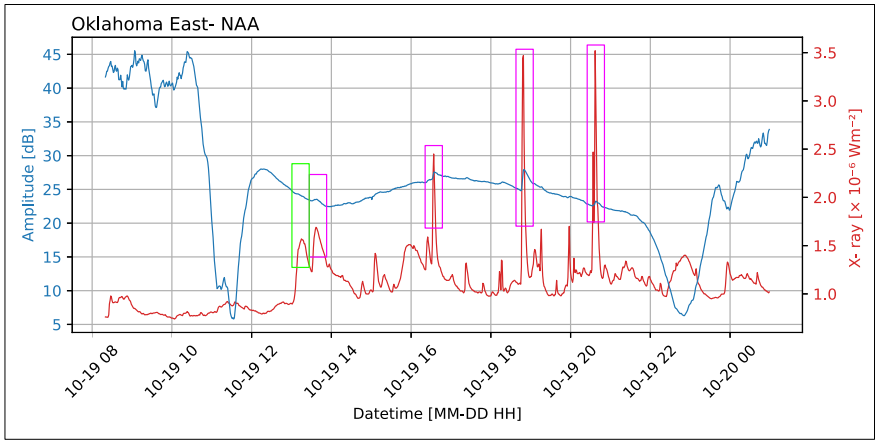

Figure 4. Example of the effects of solar flares on VLF amplitude signal variations and subsequent classification on the Oklahoma East- NAA signal; Blue line- VLF amplitude; Red line- X-ray irradiance data.

the measured values. In both instances, as well as analogous situations where a segment of the signal distinctly differs from the remainder, the signal in question may be classified as belonging to the instrumental error class. In the nighttime and solar flare classes, the topic of discussion pertained to the initiation and conclusion of the labeling process. However, in this instance, the question is clear: the initial data point that distinguishes itself from its predecessor and aligns with the visual pattern of instrumental error constitutes the first instrumental error data class, as does the final point.

2.5. Outlier data point(s)

Finally, the most uncommon data class in the VLF amplitude multi-label classifications consists of the outlier data points. Outlier data points appear to be distinctive data points that are markedly dissimilar from preceding and following data points. When analyzed alongside X-ray irradiance data, they exhibit no correlation, indicating there is no rationale for the elevated value. Figure 6 illustrates four instances in which the data points are notably dissimilar from preceding and subsequent VLF amplitude values and exhibit no correlation with X-ray irradiance data. Three of the values indicate that the outliers exhibit an increased value, while the fourth, as shown in Figure 6, demonstrates that the value can also decrease. The cause of the outlier data points remains unexplained; it may be related to internal instrument malfunctions or other factors, yet they necessitate their own data classification.

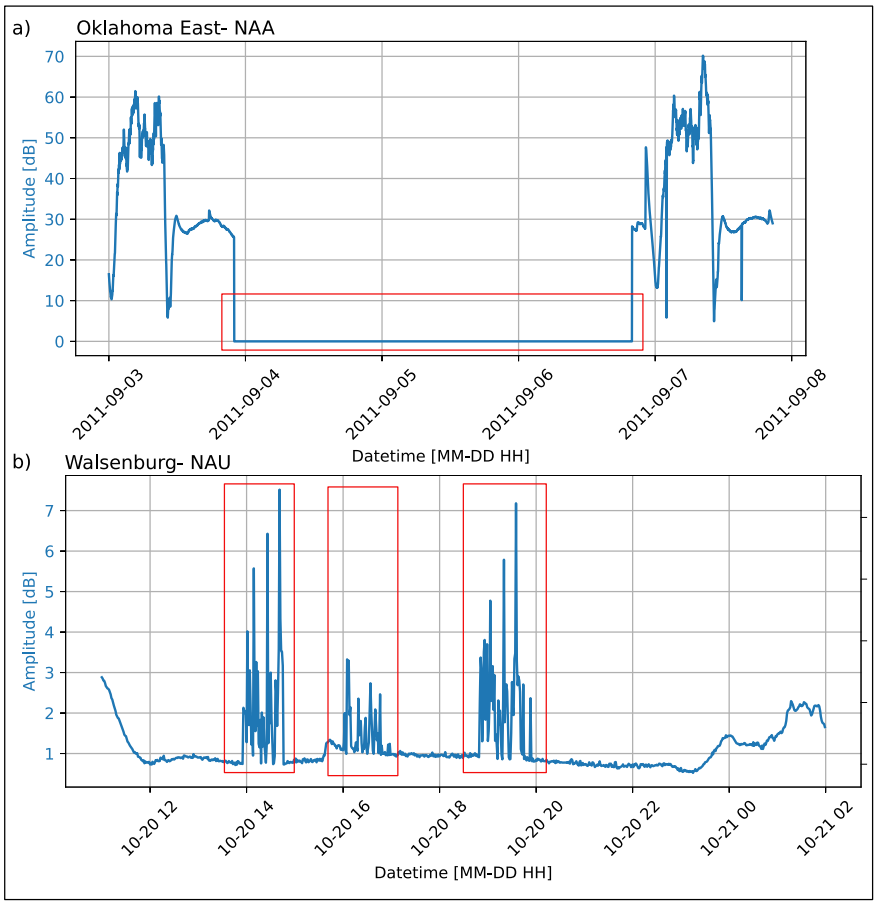

Figure 5. Types of instrumental errors on VLF amplitude signals; (a) Oklahoma East- NAA VLF amplitude signal; (b) Walsenburg- NAU VLF amplitude signal; Blue line- VLF amplitude.

3. Discussion and future prospectives

This paper addressed the standardization framework for labeling various VLF amplitude signal classes; however, the practical aspects also warrant discussion. The WALDO narrowband VLF data is provided in MATLAB data format, which is not the most versatile data format. Regarding the labeling process, rather than manually labeling each data point, our prior experience indicates that tools like TRAINSET (Geocene Inc.), a client-side, free, and open-source graphical tool for data labeling, sufficiently facilitate the laborious and time-

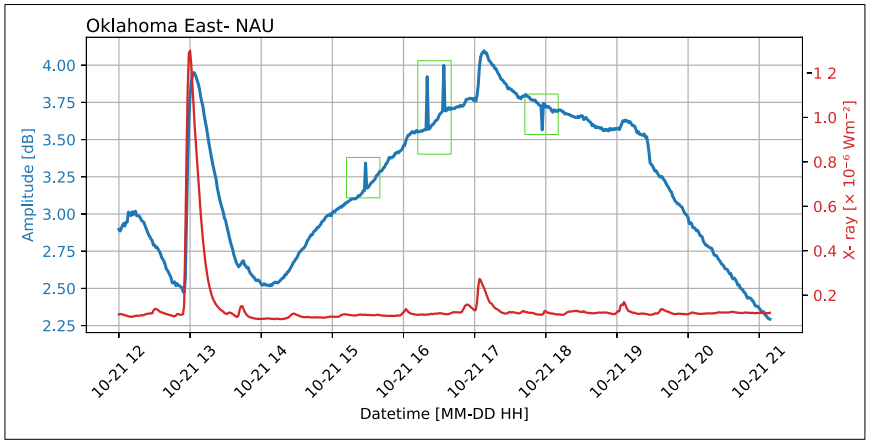

Figure 6. Examples of outlier data points on VLF amplitude signals from the Oklahoma East- NAU VLF amplitude signal; Blue line- VLF amplitude; Red line- X-ray irradiance data.

intensive task of data labeling. Given that TRAINSET requires data in a specific format and WALDO only supplied VLF amplitude and phase data, future research will focus on developing a free and open-source software package to convert MATLAB format into the designated TRAINSET format, incorporating X-ray irradiance data. Thus, data labeling will solely pertain to the labeling process itself, excluding additional issues arising from the manipulation of diverse data sources, formats, and others.

Furthermore, subsequent efforts aim to label an extensive amount of data from the WALDO database, commencing with the dataset utilized in Arnaut et al. (2023). Moreover, the multi-class ML classification can proceed with the objective of creating a model capable of autonomously identifying diverse VLF amplitude features as previously mentioned. The primary advantages of this initiative lie in data processing, as certain researchers aim to eliminate specific components of the VLF amplitude signal, including instrument errors and outliers, without manual intervention. Furthermore, if the model demonstrates adequate predictive power, a (near) real-time pipeline can be established to identify solar flare effects in (near) real-time or to address system malfunctions, among other issues. Further research is required, with the proposed standardization framework being the initial focus. The creation of specialized tools, along with the use of free and open-source resources, allows a diverse group of researchers to participate in the labeling process. Subsequent research will focus on the development of these tools, after which the data labeling process may commence.

4. Conclusion

This paper presents a standardization framework for multi-class labeling of VLF amplitude signals. The paper delineated five primary features of VLF amplitude signals, categorized as: normal (daytime) signal, nighttime signal, solar flare effects, instrumental errors, and outlier data points. A comprehensive explanation was provided for each of the five classes, including examples of the signal's appearance, the initiation of the labeling process, and its conclusion.

The practical aspect was also evaluated, including the most effective tools for this endeavor and future prospects for additional tools that will enhance the labelling process from the WALDO database to labeled data and ultimately to ML models. The primary objective of this model is to facilitate the automatic detection of all main VLF signal features, eliminating the necessity for manual labeling. Additionally, (near) real-time detection of VLF amplitude signal features is a potential outcome, although further research is required for successful implementation. The research area of automatic VLF amplitude signal features holds importance for various interdisciplinary fields related to the Sun-Earth connection, environmental studies, human health, and climate, among others and it will be the topic of more research in the future.

Acknowledgements. 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.

References

Adhikari, S., Thapa, S., & Shah, B. K., Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere. 2020, in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2-4 July 2020*

Ameer Basha, G., Lakshmana Gupta, K., & Ramakrishna, K. 2020, *Advances in Data Science and Management; Chapter: Expectation of Radar Returns from Ionosphere Using Decision Tree Technique* (Springer Nature: Singapore), DOI:10.1007/978-981-15-0978-0_20

Arnaut, F., Kolarski, A., & Srećković, V. A., Random Forest Classification and Ionospheric Response to Solar Flares: Analysis and Validation. 2023, *Universe*, **9**, DOI:10.3390/universe9100436

Arnaut, F., Kolarski, A., & Srećković, V. A., Comparative Analysis of Random Forest and XGBoost in Classifying Ionospheric Signal Disturbances During Solar Flares. 2024a, in *EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-2046*

Arnaut, F., Kolarski, A., & Srećković, V. A., Machine Learning Classification Workflow and Datasets for Ionospheric VLF Data Exclusion. 2024b, *Data*, **9**, DOI:10.3390/data9010017

Dhande, J. D. & Dandekar, D. R., PSO Based SVM as an Optimal Classifier for Classification of Radar Returns from Ionosphere. 2011, *Int. J. Emerg. Technol.*, **2**, DOI:-

Lian, J., Liu, T., & Zhou, Y., Aurora Classification in All-Sky Images via CNN–Transformer. 2023, *Universe*, **9**, DOI:10.3390/universe9050230

Oo, A. N., Classification of Radar Returns from Ionosphere Using NB-Tree and CFS. 2018, *Int. J. Trend Sci. Res. Dev.*, **2**, DOI:10.31142/ijtsrd17126

Shang, Z., Yao, Z., Liu, J., et al., Automated Classification of Auroral Images with Deep Neural Networks. 2023, *Universe*, **9**, DOI:10.3390/universe9020096

Wang, J., Huang, Q., Ma, Q., et al., Classification of VLF/LF Lightning Signals Using Sensors and Deep Learning Methods. 2020, *Sensors*, **20**, DOI:10.3390/s20041030