Nowcasting of Amplitude Ionospheric Scintillation Based on Machine Learning Techniques

OTÁVIO CARVALHO
PEDRO AUGUSTO ARAUJO DA SILVA DE ALMEIDA NAVA ALVES
Federal University of Maranhao, Sao Luis, MA, Brazil

RICARDO YVAN DE LA CRUZ CUEVA
State University of Maranhao, Sao Luis, MA, Brazil

ALEX OLIVEIRA BARRADAS FILHO
Federal University of Maranhao, Sao Luis, MA, Brazil

Ionospheric scintillation is a phenomenon that can compromise and even make the operation of some space-based systems unfeasible. In this context, it is important to develop tools capable of predicting its occurrence. However, modeling this phenomenon is quite complex due to the influence of several other aspects, such as the geomagnetic and solar activities, the seasons, and the geographic location. Therefore, the main objective of this article was to develop short-term predictive models about amplitude ionospheric scintillation through machine learning techniques. The dataset used was built considering information related to geomagnetic, solar, and interplanetary activities, the phenomenon's temporal and geographic dependence, and the ionosphere's state. To predict the value of the scintillation index S_4 30 min in advance, six models were used, based on three algorithms, the artificial neural network, the extreme gradient boosting, and the random forest. The results indicated a very satisfactory prediction capacity since a coefficient of determination of 0.87 was achieved by the lower performance model. Additionally, the results demonstrated the usefulness of the considered dataset and the feature selection approach in the model's development phase, which led to better models in accord with some statistical tests performed.

Manuscript received 20 December 2021; revised 8 April 2022; accepted 23 May 2022. Date of publication 6 July 2022; date of current version 6 December 2022.

DOI. No. 10.1109/TAES.2022.3188741

Refereeing of this contribution was handled by Z. Bi.

Authors' addresses: Otávio Carvalho, Pedro Augusto Araujo da Silva de Almeida Nava Alves, and Alex Oliveira Barradas Filho are with the Data Analytics and Artificial Intelligence Laboratory, Federal University of Maranhao, MA 65085-580, Brazil, E-mail: (otaviocarvalho.carvalho@gmail.com; pedro.alves@darti.ufma.br; alex.barradas@ecp.ufma.br); Ricardo Yvan de La Cruz Cueva is with the State University of Maranhao, MA 65055-310, Brazil, E-mail: (cueva.ryc@gmail.com). *(Corresponding author: Otávio Carvalho.)*

0018-9251 © 2022 IEEE

I. INTRODUCTION

The term ionospheric scintillation refers to the occurrence of variations in parameters, such as amplitude, phase, propagation direction, and polarization of waves that propagate in the ionosphere. Such variations are caused by fluctuations in the refractive index of the ionospheric layer, related to irregularities in the medium's electronic density [1], [2]. This phenomenon is the source of both scientific and practical interests since it affects, among other systems, a series of space-based applications, which includes launch vehicles' operation, military telecommunications, and satellite navigation systems [3]–[6].

Considering these aspects, the development of predictive techniques, both short and long-term, aimed at predicting ionospheric scintillation becomes an increasingly urgent need [7]. However, building such models is not a simple task, given the complexity of this phenomenon [8]. Several factors affect the occurrence and intensity of ionospheric scintillation, among them, it is worth mentioning the geomagnetic, solar, and interplanetary activities, local time, geographic location, in addition to the spatial and temporal characteristics of the ionosphere [9]–[11]. Amid the various initiatives intended at developing predictive models focusing on ionospheric scintillation, the use of machine learning techniques is a growing approach.

In [12], the authors used machine learning techniques known as the bagging method and the classification and regression decision tree (CART) algorithm to predict the amplitude scintillation index S_4 with advances of 1–4 h. In addition, they also performed a qualitative prediction of the level of ionospheric scintillation the next day among the possibilities of weak, moderate, or strong scintillation. A particularity of this article was the use of information collected in different cities. In [8], the authors applied the CART algorithm to assess the correlation between the daily occurrences of scintillation in two Brazilian cities. Their results showed a certain degree of correlation among the scintillation occurrence in the different cities and provided a direction for the development of improved predictors.

In [9], the authors developed models that could predict the S_4 index 1–3 h in advance. The models were built through the artificial neural network (ANN) algorithm and used data collected in Australia. Additionally, the authors conducted a correlation study to aid in the model development phase. In [13], the authors used the ANN algorithm to develop models that could predict the S_4 index a day in advance using data collected at Guam (13.58°E, 144.86°N). A particularity of this article was the use of a genetic algorithm together with the machine learning approach.

In [11], the authors applied the support vector machine (SVM) algorithm to develop predictive models related to amplitude scintillation in the equatorial region. The authors used signal-derived parameters as inputs to the machine learning model to identify scintillation events. In [14], the authors employed three gradient boosting-based algorithms to identify the occurrence of strong scintillation events in the Brazilian longitude sector. The authors explored data from several sources, aiming to represent some of the underlying phenomena that affect the scintillation's dynamics. The information used incorporated solar wind, ionosonde, geomagnetic index, and local ionospheric data. In [15], the authors used the decision tree algorithm to identify scintillation events using features collected by a receiver designed to study the scintillation phenomenon. In [16], the authors applied the machine learning algorithm extreme gradient boosting (XGBoost) to develop classification models regarding the identification of amplitude scintillation events. Their results showed that the proposed approach did better than other well-known techniques.

The goal of this article is the development of models regarding the short-term prediction of the S_4 amplitude scintillation index, 30 min in advance, through machine learning techniques. The S_4 index is defined as the standard deviation of the signals' intensity normalized by its mean value and is used to quantify the amplitude scintillation [15]. The models are built considering two Brazilian cities, São Luís (2.3° S, 44.2° W) and Cachoeira Paulista (22.7° S, 45.0° W). These locations were selected as a result of some special conditions found there. First, the Brazilian territory is located in one of the regions most affected by ionospheric scintillation from a global perspective [17]. In addition, São Luís is located near the magnetic equator, a region characterized by the formation of large-scale structures of depleted plasma density in the ionosphere ranging from hundreds of kilometers in the east–west magnetic direction, known as equatorial plasma bubbles (EPB). These bubbles extend for thousands of kilometers in the magnetic north–south direction and are responsible for the scintillation of radio signals that propagate through them [8]. Besides that, Cachoeira Paulista is located under the crest of the equatorial ionization anomaly (EIA), an area highly affected by ionospheric scintillation [18].

In the present report, the models developed are aimed at the S_4 index's prediction at the city of Cachoeira Paulista, using information collected at São Luís, exploring the ionospheric scintillation's regional dependence. The study is based on the results presented in [8] and [12], where the authors demonstrate the relationship between the scintillation's occurrence at the magnetic equator and in the EIA's region through the EPB's influence.

The contributions and novelty of this study lie in the development of a nowcasting model for the S_4 index, something still little explored through machine learning techniques, since most of the work focuses on prediction horizons in the order of hours. Besides the construction and analysis of an extensive dataset, comprising several phenomena related to ionospheric scintillation. The underlying aspects considered are the geomagnetic, solar, and interplanetary activities, the phenomenon's time dependence, and the background of the ionosphere. The geomagnetic activity is represented by data collected through magnetometers, the solar and interplanetary activities are characterized by data from solar wind sensors, the phenomenon's time dependence is modeled considering information related to the local time, and the ionospheric background is represented by data from global positioning system (GPS) sensors and ionosondes. Particularly, to the best of our knowledge, the use of the magnetometer's data and the time-dependence modeling used were not employed previously to predict the S_4 index by machine learning techniques. Through this analysis, it will be possible to examine how such parameters affect the models developed, indicating subsequent steps in the direction of better models' development. Furthermore, some feature selection approaches are used in the machine learning models' construction phase, intending to assess how such methodologies influence the performance of the developed predictive models.

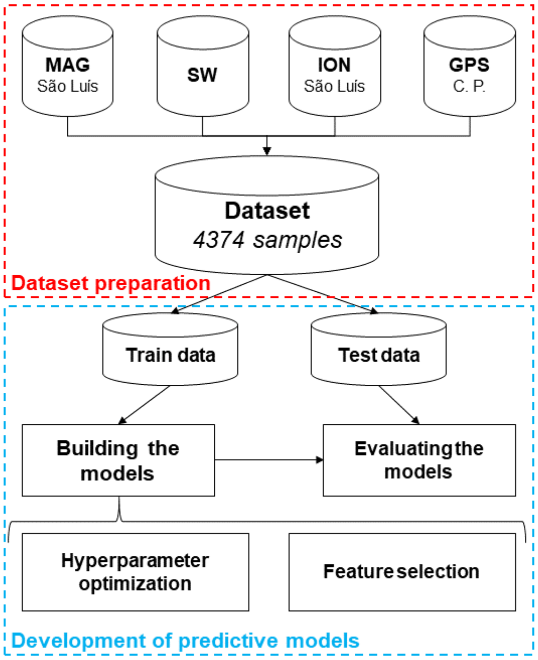


Fig. 1. Applied methodology.

This rest of this article is organized as follows. In Section II, the dataset construction process and the machine learning models' development are described. Section III presents the study results and their implications. Finally, Section IV conludes this article.

II. METHODOLOGY

In this section, the necessary steps to reach the results will be described. Fig. 1 illustrates the process conducted in this study, organized in two main phases explained in the next subsections: the *Dataset preparation* and the *Development of predictive models*.

A. Dataset Preparation

First of all, the period considered to collect data comprised the months of October and November since the seasonal variation of the phenomenon studied leads to an intensification of its effects in the Brazilian territory from September to March [19], [20]. Furthermore, the year 2014 was selected because it corresponds to a great level of activity in solar cycle 24, which intensifies the ionospheric

TABLE I
Features Related to Geomagnetic Activity

Feature	Description
F_M	Absolute Value of the main field
H_M	Horizontal Component of the magnetic field vector
D_M	Magnetic Field Declination

TABLE II
Features Related to the Solar and Interplanetary Activities

Feature	Description
IMF	Mean Magnitude of the interplanetary magnetic field
B_z	z Component of the interplanetary magnetic field
P_P	Plasma Flow Pressure
IEF	Interplanetary Magnetic Field

TABLE III
Features Related to the State of the Ionosphere Derived From Ionosonde

Feature	Description
foF2	Critic Frequency of the F2 Layer
fminF	Minimum Frequency of the layer F echoes
fmin	Minimum frequency of the ionogram echoes
h'F2	Minimum virtual height of the F2 trace
h'F	Minimum virtual height of the F trace
fxI	Maximum frequency of the F-trace
FF	Frequency Distribution between *fxF2* and *fxI*
hmF2	Peak height of the F layer
yF2	Half the thickness of the F2 layer, parabolic model
TEC	Total Electron Content
f(h'F)	Frequency at which *h'F* occurs
f(h'F2)	Frequency at which *h'F2* occurs

TABLE IV
Features Related to the Background of the Ionosphere Derived From a GPS Receiver

Feature	Description
ROTI	ROTI Index
θ_E	Satellite Elevation Angle
θ_A	Satellite Azimuth Angle
MA-S4	S_4 Index value at the current time

scintillation [21]. Fig. 1 illustrates the dataset construction process at the dashed red rectangle. It was considered information regarding the geomagnetic activity, indicated by the MAG dataset, the solar and interplanetary conditions, indicated by the *SW* dataset, and the background of the ionosphere, indicated by the ION and GPS datasets. In addition, Fig. 1 indicates the cities from which the information was collected, where *C.P.* means Cachoeira Paulista. The indicated datasets will be explored individually as follows.

To build the MAG dataset, data were collected by magnetometers, devices employed to measure the intensity and direction of the Earth's magnetic field [22]. Such information was made available through the portal *Space Weather Data Share*,[1] maintained by the Brazilian Space Weather Study and Monitoring program, Embrace/INPE.[2] The sensors used originate from the MagNet network, consisting of saturated core-type magnetometers manufactured by the Radio Observatory of Jicamarca, sensitive to magnetic fields of the order of 1 mT, with a maximum resolution of up to 10 pT [23]. The three parameters considered are arranged in Table I.

The *SW* dataset was represented using data from the *OMNIWeb* tool,[3] a database maintained by NASA that gathers information, collected by several orbiting satellites, about the magnetic field and plasma that constitute the solar winds [24]. The four parameters from this source used in this study are indicated in Table II.

Ionosonde information was used to build the ION dataset. Ionosondes represent an important tool for both research and monitoring operations of the ionosphere [25]. The data associated with this instrument were obtained through the *Space Weather Data Share* portal, similarly to the data related to magnetometers. The ionosonde responsible for data acquisition belongs to the digital ionosonde (DGS) 256 category [26]. The quality of information extracted from these devices can be measured on a scale from 0 to 100 by a parameter known as *confidence score* (CS). This metric relies on quality criteria related to the uncertainty and confidence of the results derived from ionosondes. In this study, only samples with a CS greater than 40 were considered, a criterion to ensure the quality of the data used [27]. The twelve parameters associated with the ionosonde used in this study are listed in Table III.

The ionosonde parameters listed in Table III give information about the ionospheric background conditions, which can be directly related to the EPB generation [28]. Considering the EPB is important because they cause rapid fluctuations in amplitude and phase of radio signals, producing severe ionospheric scintillation in radio waves [29].

Data from GPS receivers are widely used by the scientific community for studies focused on the dynamics of the ionosphere and, particularly, are extensively used in research focusing on ionospheric scintillation [30]. The four parameters used in this study associated with these devices are listed in Table IV.

At this point, an important aspect should be addressed. The GPS sensor whose data were acquired collects data from several satellites at any timestamp. Thus a proper strategy is needed to select a suitable satellite to retrieve data. In this situation, it was chosen to apply a methodology inspired

[1] [Online]. Available: http://www2.inpe.br/climaespacial/SpaceWeatherDataShare/

[2] [Online]. Available: http://www2.inpe.br/climaespacial/portal/en/

[3] [Online]. Available: https://omniweb.gsfc.nasa.gov/index.html

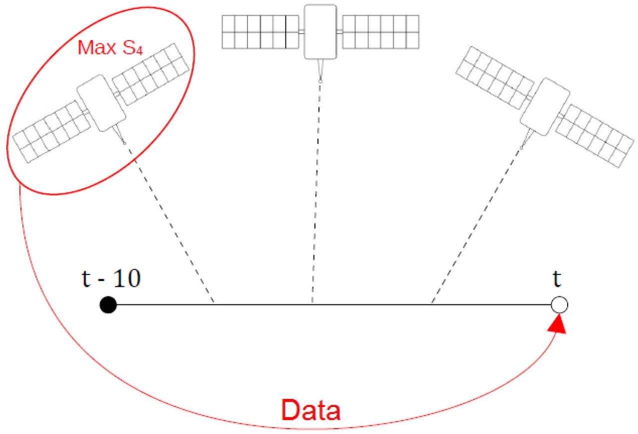
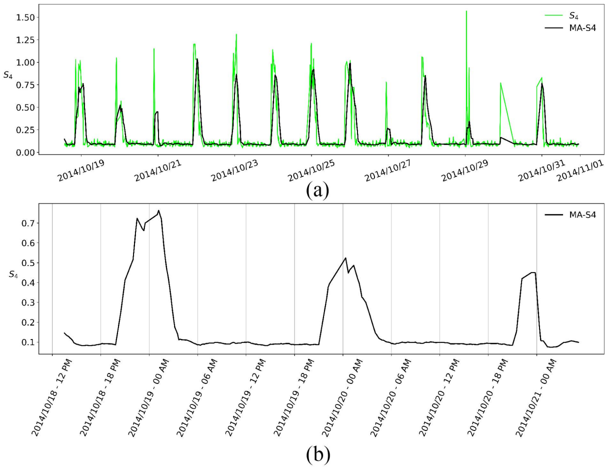


Fig. 2. GPS dataset data processing.



(a)

(b)

Fig. 3. Results obtained by processing GPS dataset data.

TABLE V
Features Related to the Time Dependence of the Ionospheric Scintillation

Feature	Description
DNS	Sine component of day number of the year
DNC	Cosine component of day number of the year
HRS	Sine component of hour number of the day
HRC	Cosine component of hour number of the day

by the approach used in [12]. The processing occurred in two sequential steps, described as follows:

1) Consider only satellites whose elevation angle is greater than 30°, that is, $\theta_E > 30^\circ$. This step is necessary to mitigate the effects of multipath, a phenomenon that negatively affects the signals captured by GPS receivers [31]
2) For a given instant of time *t*, verify, in 10 min prior to that instant, which satellite presents the maximum value of the S_4 index. Then, select the data from this satellite as representative of the instant *t*.

Fig. 2 presents an illustration of the steps described above. In this processing step, it was possible to obtain a time series of the S_4 index. Intending to smooth the curve that describes the variation of this parameter, as in [12] a moving average operation was applied, generating the *MA-S4* parameter. This operation comprised 15 samples, a number determined by inspection, seeking a compromise between smoothing the data's noise and keeping its original variation characteristics. The objective of the predictive models developed in this article is to perform the prediction of the *MA-S4* parameter 30 min in advance.

Fig. 3 shows the results of the GPS dataset processing step. In Fig. 3(a), the green curve represents the time series of the S_4 index obtained through the processing steps indicated above; the black curve indicates the result of the performed moving average operation. Clearly, the moving average operation was efficient in terms of reducing the amount of noise present in the data. In Fig. 3(b), it is possible to analyze a zoom of the initial region of the graph shown in Fig. 3(a), which indicates a periodic behavior in the variation of the S_4 index characterized by an increase that starts around 6 P.M. and extends to midnight, where the highest intensities of the S_4 index are observed. Sometimes, this behavior extends for a few hours longer and decays hours later. This characteristic pattern is observed in operating signals in the L frequency band in the equatorial ionosphere region according to [7] and [32]. This observation indicates that the preprocessing steps illustrated in Fig. 2 maintained the characteristics inherent to the ionospheric scintillation phenomenon, and therefore, are adequate for the analysis developed in this study.

In addition to the aspects mentioned above, the temporal dependence of the phenomenon was considered. To consider the annual and daily variations of ionospheric scintillation, the parameters indicated in Table V were used, obtained through the following equations [33]:

$$\mathrm{DNS} = \sin\left(\frac{2\pi \cdot D}{365.25}\right) \tag{1}$$

$$\mathrm{DNC} = \cos\left(\frac{2\pi \cdot D}{365.25}\right) \tag{2}$$

$$\mathrm{HRS} = \sin\left(\frac{2\pi \cdot H}{24}\right) \tag{3}$$

$$\mathrm{HRC} = \cos\left(\frac{2\pi \cdot H}{24}\right). \tag{4}$$

In (1) and (2), the term D denotes the number of the considered day, therefore, $1 \leq D \leq 365$. In (3) and (4), the term H refers to the number of the considered hour, hence, $0 \leq H \leq 23$ [33]. Both variables D and H were measured on the local time (LT) scale (LT = UT - 3). The resulting dataset has 4374 rows, referred to samples, and 27 columns. Furthermore, the dataset was standardized to have a zero mean and unitary variance.

B. Development of Predictive Models

Fig. 1 illustrates the predictive models' development process in the dashed blue rectangle. At this point, the first step consisted of splitting the dataset into two subsets, Train and Test data. The approximately sixty days that make up the original dataset were divided into four bins of approximately fifteen days. The first three bins (3486 samples) were used to build the models (Train data), and the last one (888 samples) was held out to be used only for testing them (Test data). Additionally, some terms related to machine learning will be defined. In machine learning, training a model means providing samples from the dataset so the algorithm used can detect the existing patterns. The model's input parameters are also called features. Thus the dataset used in this study comprises 27 features and one variable of interest, also called output or y.

To obtain the predictive models in this study, the algorithms *random forest* (RF), *artificial neural network* (ANN), and *extreme gradient boosting-XGBoost* (XGB) were used. The RF is an ensemble model, which implies its predictions result from a combination of various submodel answers. In this case, the submodels are tree predictors, each one related to the values of a random vector sampled independently and with a distribution common to the whole forest of trees [15]. The ANN algorithm provides a way to build complex models by the combination of simple processing units, called neurons, in a layered arrangement. In general, the gradient descent method is used to train an ANN model [34]. The XGB algorithm produces a scalable model for tree boosting. Likewise the RF algorithm, the boosting approach uses an ensemble to make predictions, however, it differs from the former as long as boosting builds the model sequentially, with new submodels using information from the previous ones to improve their performances [35], [36].

1) *Hyperparameter Optimization:* Hyperparameter optimization comprises the process of defining a set of hyperparameters suitable for the developed predictive model, a decisive step in the accuracy of the obtained results [37], [38]. In this article, we used a standard tool, known as OPTUNA, to optimize the model's hyperparameters. OPTUNA is a tool that approaches the hyperparameter optimization process as a task of minimization/maximization. It uses an objective function that takes the hyperparameters as inputs and returns a score used to select the best combination of hyperparameters. Combining efficient searching and pruning algorithms to improve the optimization process, OPTUNA's scalable and versatile design allows for its use in a broad range of applications [39].

2) *Feature Selection:* This step consists of selecting a subset of representative features from the dataset, which allows the construction of satisfactory models and at the same time reduces undesirable effects, such as noise and computational cost associated with unnecessary features [40]. In this article, the feature selection approach will be the filter strategy, which consists of determining a features' importance ranking and using only the most relevant ones [41]. The features' importance was evaluated based on the mutual information criterion and the use of an RF model. The mutual information criterion is seen as a satisfactory way to assess the degree of relationship between datasets since it can detect dependencies of any type between the considered variables [42], [43]. The approach related to the RF algorithm explores the fact that this technique offers a way to estimate the importance of different features during the training stage [44]. In Table VI it is possible to observe the feature importance's rankings obtained through the mutual information criterion, R_1, and the RF model, R_2.

The next step in feature selection requires determining the optimal set of features to be employed. This process was performed as follows: An iterative process was executed in which in the jth iteration a model was trained considering the best j features according to one of the ranking processes described. This model was then evaluated using some evaluation metrics. At the end of the 27 iterations, the optimal number of features is defined as that associated with the best-performing model, determined through the coefficient determination metric (R^2), described in (5). For each feature importance ranking, this methodology was applied considering the ANN, XGB, and RF algorithms, resulting in six predictive models, two for each algorithm.

In the model training step, the cross-validation technique k-fold [45] was used. However, because the variable of interest and the features used are in the form of a time series, which have some internal dependence on the data, characterized by the tendency of close observations to present greater similarity, it is necessary to modify the cross-validation strategy [46], [47]. This aspect was addressed using an approach that consists of dividing the dataset into groups of ordered samples, which are then distributed among the training and validation sets used in the k-fold methodology [46], [48].

In Fig. 4, the quantities, as well as the features selected by the approach explained previously, are presented. The labels on the left side of Fig. 4 indicate the algorithm and the feature importance ranking used to select the features. For instance, the label ANN_1 represents the model built using the ANN algorithm together with the ranking R_1 of feature importance. The number of selected features is indicated on the right side. In Fig. 4 the black squares indicate the selected features.

3) *Evaluating the Models:* To check the models' performance the evaluation metrics used were the coefficient of determination (R^2), the mean squared error (mse), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). The expressions used in the determination of these metrics can be conferred in the following equations:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{5}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{6}$$

TABLE VI
Rankings of Feature Importance

Ranking Technique	Feature importance ranking
R_1	*MA-S4 – CCH – CSD – CCD – yF2 – CSH – TEC – fxI – foF2 –* H_M *– ROTI –* F_M *–* θ_A *– fminF – IMF – f(h'F2) – hmF2 – f(h'F) – FF –* D_M *–* P_P *– h'F2 – h'F – fmin –* θ_E *– IEF –* B_z
R_2	*MA-S4 – ROTI – CSH – h'F2 – h'F – hmF2 –* P_P *–* θ_A *– yF2 – IMF –* F_M *– fxI –* H_M *– IEF –* θ_E *– FF – TEC – CSD – CCD –* B_z *–* D_M *– foF2 – fmin – f(h'F) – f(h'F2) – fminF – CCH*

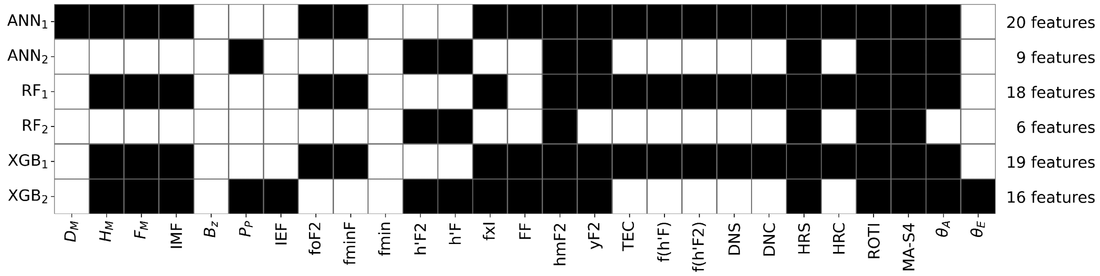


Fig. 4. Features used in each model.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{7}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N} \left|\frac{\hat{y}_i - y_i}{y_i}\right|. \tag{8}$$

In (5)–(8): $\hat{y}$, y, and $\bar{y}$ denote, respectively, the predicted value, the actual value, and the average value of the variable y. Furthermore, N represents the number of samples used. In addition to computing evaluation metrics to compare different machine learning models, it is necessary to assess whether the observed differences in the metrics' values are statistically relevant. A stage in which statistical significance tests are usually performed [49]. Under these circumstances, the most suitable test is the Wilcoxon Signed Rank test [50], which was adopted in this study. The necessary computational implementation was performed largely through the Python libraries *Scikit Learn* [51], and *keras* [52].

III. RESULTS AND DISCUSSION

In this section, the results are organized into two parts. Section III-A focuses on the results related to the dataset and the feature selection strategies, and the Section III-B presents the results associated with the predictive models.

A. Dataset and Feature Selection

Analyzing the results arranged in Fig. 4, the feature *MA-S4* displays at the top of the rankings provided by feature selection algorithms, being present in all the sets of the selected features. It reinforces the fact that models aimed at predicting ionosphere characteristics have their performance significantly improved through the use of features containing information delayed in time, as found in [53]. Also, the feature ROTI was selected in all feature sets concerning the two machine learning algorithms, possibly due to this index's ability to capture effects associated with ionospheric irregularities that occur on the minute scale and by its correlation with the S_4 index in the generation, evolution, and decay phases of the ionospheric scintillation phenomenon [25], [54].

Additionally, the presence of several parameters from ionosondes in the selected feature sets shows the benefit of such data in studies related to ionospheric scintillation. This could be explained by the relation between ionosonde data and the EPB's dynamics, which is directly related to the occurrence of ionospheric scintillation [28], [29]. Since it is known that the critical frequencies of the ionospheric layer, quantities well-defined and regularly measured by ionosondes, constitute some of the most widespread parameters to describe the state of the ionospheric layer, which permits associating the frequency *foF2* directly to the occurrence of scintillation [55], [56]. In addition to the critical frequencies, the heights of the different regions of the ionosphere are useful to characterize the state of the layer, such as the parameters $h'F$ and $hmF2$, which enable the analysis of the vertical movements of the ionospheric plasma in the F layer, as seen in [57] and [58]. Additionally, the peak height of the F layer ($hmF2$) is present in all models derived from the feature selection strategy, as depicted in Fig. 4. It is expected since at the heights of peak electron density takes place most of the scintillation-producing irregularities. Thus, the changes in the parameter $hmF2$ influence the dynamic of these irregularities [59]. This result agrees with [14], where the authors found that the parameter $hmF2$ plays an important role in the scintillation phenomenon.

Another aspect to be highlighted concerns the selection of time-related features. In Fig. 4, it is possible to verify that in all models at least three time-related features were present, which reinforces the usefulness of considering the time dependence of ionospheric scintillation in predictive models. Also, the features referring to solar winds were selected a few times in comparison with the parameters

TABLE VII
Results Obtained Using the ANN Algorithm

Model	R^2	MSE	MAE	MAPE
ANN_0	0.9612 ± 0.0287	0.0009 ± 0.0007	0.0154 ± 0.0053	0.8729 ± 0.1039
ANN_1	0.9512 ± 0.0179	0.0011 ± 0.0004	0.0199 ± 0.0072	0.9315 ± 0.1866
ANN_2	0.9533 ± 0.0155	0.0011 ± 0.0004	0.0183 ± 0.0051	0.8744 ± 0.1121

related to the other phenomena considered. It may be related to the fact that the solar wind has a greater impact on the incidence of ionospheric scintillation in the polar regions, as argued in [60]. The SW feature more selected was the IMF, which could be due to the association of this parameter to the occurrence of magnetic storms [61]. This indicates that, eventually, it may be more appropriate to seek other representative parameters of solar activity and interplanetary environment for use in ionospheric models aimed at the equatorial region. To represent the influence of solar activity in the equatorial region the F10.7 index can be used, as shown in [29], where the author found a relation between it and the S_4 index. To represent the interplanetary conditions, coronal mass ejection (CME) information can be used as it influences the dynamics of scintillation [62]. Also, the selection of variables acquired through magnetometers is noteworthy, what is associated with the influence of geomagnetic effects in ionospheric scintillation, as verified in [25] and [63].

Concerning the predictive models, analyzing Fig. 4, it is possible to note that even from the same ranking, the number of features selected was different for each algorithm. It emphasizes that different sets of features have distinct importance when considering different algorithms. Furthermore, the ranking strategies led to different results, as shown in Table VI. It occurred because the methodologies adopted in this stage have different ways of quantifying the relationship between the features and the variable of interest. The approach based on mutual information, theoretically, can identify more complex dependencies between the variables [42], [43]. However, the method that uses the RF algorithm, according to [64], can evaluate subsets of features instead of considering them individually during the training process.

B. Predictive Models

In this section, the quantitative results from the developed models will be described. Nine different models were built, three for each of the algorithms used, defined as follows, one model considering all available features (indicated by the 0 subscript) and the remaining two using the optimal sets of features determined through the methodology described in Section III-B from the rankings R_1 and R_2 (indicated by the subscripts 1 and 2, respectively). The results obtained through the training set concerning the ANN, XGB, and RF algorithms are summarized, respectively, in Tables VII–IX through the mean and the standard deviation of the considered evaluation metrics.

In Tables VII and VIII it is not possible to indicate a model superior to the others just by analyzing the metric values. In Table IX, considering just the metric values, it seems that the RF_2 model is a bit better than the other RF models. Additionally to the computation of evaluation metrics, some statistical tests were applied to compare the performances of different models. In all these tests, the R^2 metric is used to assess the models' performance.

The approach used consisted in comparing pairs of models. The first set of tests comprised the models related to the ANN and XGB algorithms. Concerning the models ANN_0 and XGB_0, the data used provided enough evidence to accept the alternative hypothesis that the former model had the best performance. Additionally, the tests related to the models ANN_1 and XGB_1, as well as the ones associated with the models ANN_2 and XGB_2, showed that the data used do not provide evidence to refute the null hypotheses of equivalence between them.

The hypothesis tests regarding the models related to the ANN and RF algorithms presented the following results: The models ANN_0 and RF_0 have equivalent performances, and the models RF_1 and RF_2 did better than the models ANN_1 and ANN_2, respectively. Concerning the comparison between the models related to the XGB and RF algorithms, the data provided sufficient evidence to accept the alternative hypotheses that the RF models outperform the XGB ones in all analyzed cases.

About the comparisons between the models derived from the same algorithm, the hypothesis tests showed that the data provided sufficient evidence to accept the alternative hypothesis that the models with selected features outperform the ones that used all the features available in two of the three algorithms used, the RF and the XGB. Only for the models derived from the ANN algorithm, did the model that used all features surpassed the ones derived from a reduced set of selected features. In summary, the features selection strategy led to improved results in the models derived from two of the three algorithms evaluated, namely, the RF and the XGB ones.

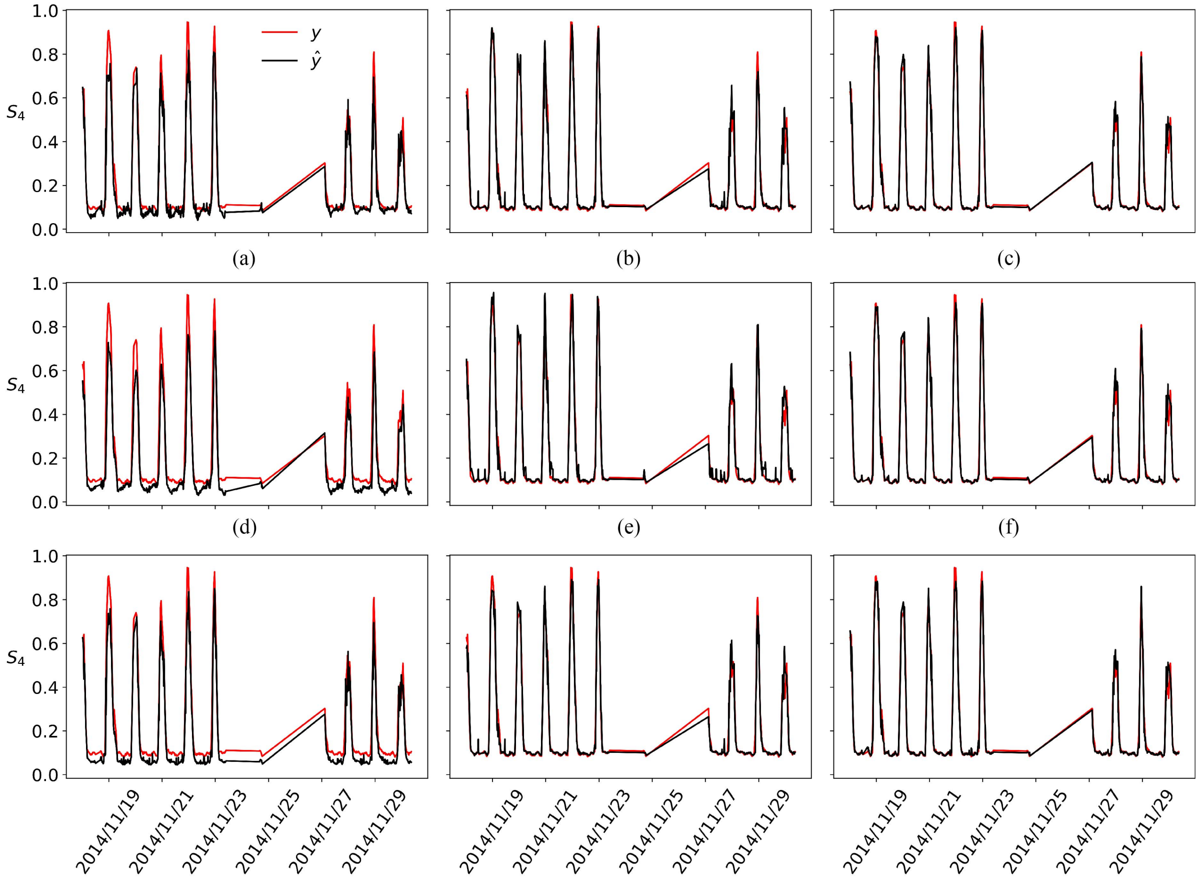
To sum up the results presented so far, the feature selection approach proved to be advantageous, since the models obtained through this strategy presented equivalent or even better performances than those that used all features, which contributed to reducing the computational cost of the models. Furthermore, the analysis regarding the feature importance indicates which information is more relevant for the construction of models aimed at ionospheric scintillation.

TABLE VIII
Results Obtained Using the XGB Algorithm

Model	R^2	MSE	MAE	MAPE
XGB_0	0.9443 ± 0.0176	0.0013 ± 0.0005	0.0172 ± 0.0043	0.8576 ± 0.1020
XGB_1	0.9550 ± 0.1020	0.0010 ± 0.0002	0.0140 ± 0.0019	0.8382 ± 0.0711
XGB_2	0.9538 ± 0.0110	0.0010 ± 0.0003	0.0141 ± 0.0020	0.8364 ± 0.0653

TABLE IX
Results Obtained Using the RF Algorithm

Model	R^2	MSE	MAE	MAPE
RF_0	0.9612 ± 0.0097	0.0009 ± 0.0003	0.0127 ± 0.0022	0.8268 ± 0.0496
RF_1	0.9621 ± 0.0087	0.0009 ± 0.0002	0.0121 ± 0.0017	0.8311 ± 0.0542
RF_2	0.9627 ± 0.0078	0.0008 ± 0.0002	0.0120 ± 0.0016	0.8266 ± 0.0466



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 5. Results obtained from the test set. (a) ANN_0 Model. (b) XGB_0 Model. (c) RF_0 Model. (d) ANN_1 Model. (e) XGB_1 Model. (f) RF_1 Model. (g) ANN_2 Model. (h) XGB_2 Model. (i) RF_2 Model.

In Fig. 5 we present the performances of the models in comparison with the real data of the Test dataset. The label y refers to the real value of the S_4 index while $\hat{y}$ indicates the model's predicted value. The plots in the left panel refer to the ANN models, the ones in the center to the XGB ones, and those on the right to the RF models. The plots (a)–(c) contain the results of the models that used all the available features, the plots (d)–(f) include those obtained from the R_1 ranking, while the plots (g)–(i) correspond to the models derived from the R_2 ranking. It appears that the RF models seem to have achieved better performances than the ANN and XGB models on the test set. In addition, the performance of the models in general improved after the feature selection for the ones derived from the XGB and RF algorithms, which

TABLE X
Evaluation Metrics Derived From the Test Data

Model	R^2	MSE	MAE	MAPE
ANN_0	0.9289	0.0023	0.0304	0.1814
ANN_1	0.8687	0.0043	0.0460	0.3058
ANN_2	0.9055	0.0031	0.0445	0.3270
XGB_0	0.9711	0.0009	0.0147	0.0702
XGB_1	0.9640	0.0012	0.0182	0.0952
XGB_2	0.9715	0.0009	0.0145	0.0639
RF_0	0.9781	0.0007	0.0127	0.0549
RF_1	0.9754	0.0008	0.0134	0.0576
RF_2	0.9776	0.0007	0.0133	0.0596

may indicate that this step was important for the removal of variables that negatively affected the developed models. Note that the gaps presented in the graphics are caused by the nonavailability of data in that period. A more complete report on the models' performance concerning the Test data can be found in Table X.

Fig. 5 also provides a visual indication that the ANN and XGB models presented a lower performance than the RF models with regard to the prediction of low levels of the S_4 index, while the RF models obtained a satisfactory performance in all magnitudes of the predicted variable. The relatively poor performance of the ANN models in the low levels of the S_4 index could indicate that weak scintillation events are more difficult to be predicted than the stronger cases, in accord with the results presented in [14]. The better performance of the XGB models regarding the ANN-based ones could be related to the algorithms' characteristics since ANN models are more suited for large datasets while gradient boosting models perform better in small ones [65]. Considering the results presented so far, it is possible to indicate that the models related to the RF algorithm presented the best results in comparison with the ones derived through the ANN and XGB algorithm.

IV. CONCLUSION

This article presented the development of models aimed at short-term prediction of ionospheric amplitude scintillation in Brazil through machine learning techniques. It was possible to predict the value of the scintillation index S_4 30 min in advance through different machine learning algorithms with very good prediction accuracy in general. The main contributions of this study are related to the construction and analysis of an extensive dataset comprising features associated with underlying phenomena that influence ionospheric scintillation. In addition, it was shown that the feature selection approach could be useful to build better machine learning models aimed at the ionospheric scintillation issue. Since it was possible to build equivalent and even better models through this strategy when compared with the models that used all the available features. The limitations of this article refer to the existence of gaps in the datasets used, mainly due to the existence of periods in which there was no available information. Another shortcoming related to the dataset is the difficulties with the real-time processing of the solar wind data, which can be computationally expensive. Thus a natural development of this article could focus on this aspect. Furthermore, the developed models considered only two cities and a reduced time interval, aspects that can be explored in future work through the inclusion of new data and the consideration of longer time intervals. Another improvement can be made by considering other feature selection strategies since the ones used do not consider the internal relationships between the features, which can lead to the use of redundant information in the models. Finally, it is possible to conclude that the developed models further demonstrate the potential of using machine learning techniques in studies aimed at ionospheric scintillation. The results presented in this report can be used to develop better predictive models related to ionospheric scintillation through the features analyzed and the proposed methodology.

ACKNOWLEDGMENT

The authors would like to thank professors E. Rodrigues de Paula and M. Banik de Pádua for their assistance in the data collection stage of the work. They would also like to thank the MCTI EMBRACE/INPE Program for providing data for this article. The authors acknowledge the use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service, and OMNI data.

REFERENCES

[1] S. Priyadarshi, "A review of ionospheric scintillation models," *Surv. Geophys.*, vol. 36, no. 2, pp. 295–324, 2015.

[2] J. Vilà-Valls, P. Closas, C. Fernández-Prades, and J. T. Curran, "On the mitigation of ionospheric scintillation in advanced GNSS receivers," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1692–1708, Aug. 2018.

[3] ESA, Paris, France, "Space weather study results," 2016. [Online]. Available: https://esamultimedia.esa.int/docs/business_with_esa/Space_Weather_Cost_Benefit_Analysis_ESA_2016.pdf

[4] M. A. Kelly, J. M. Comberiate, E. S. Miller, and L. J. Paxton, "Progress toward forecasting of space weather effects on UHF satcom after operation anaconda," *Space Weather*, vol. 12, no. 10, pp. 601–611, 2014.

[5] T. E. Humphreys, M. L. Psiaki, B. M. Ledvina, A. P. Cerruti, and P. M. Kintner, "Data-driven testbed for evaluating GPS carrier tracking loops in ionospheric scintillation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 4, pp. 1609–1623, Oct. 2010.

[6] Y. Jiao, D. Xu, C. L. Rino, Y. T. Morton, and C. S. Carrano, "A multifrequency GPS signal strong equatorial ionospheric scintillation simulator: Algorithm, performance, and characterization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1947–1965, Aug. 2018.

[7] S. Basu, K. Groves, S. Basu, and P. Sultan, "Specification and forecasting of scintillations in communication/navigation links: Current status and future plans," *J. Atmospheric Solar-Terrestrial Phys.*, vol. 64, no. 16, pp. 1745–1754, 2002.

[8] G. de Lima et al., "Correlation analysis between the occurrence of ionospheric scintillation at the magnetic equator and at the southern peak of the equatorial ionization anomaly," *Space Weather*, vol. 12, no. 6, pp. 406–416, 2014.

[9] M. Sridhar, D. V. Ratnam, K. P. Raju, D. S. Praharsha, and K. Saathvika, "Ionospheric scintillation forecasting model based on NN-PSO technique," *Astrophys. Space Sci.*, vol. 362, no. 9, 2017, Art. no. 166.

[10] S. Priyadarshi, "Ionospheric scintillation modeling needs and tricks," in *Proc. Satellites Missions Technol. Geosciences*, 2020, pp. 1–14.

[11] Y. Jiao, J. J. Hall, and Y. T. Morton, "Automatic equatorial GPS amplitude scintillation detection using a machine learning algorithm," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 1, pp. 405–418, Feb. 2017.

[12] L. F. C. Rezende et al., "Survey and prediction of the ionospheric scintillation using data mining techniques," *Space Weather*, vol. 8, no. 6, pp. 1–10, 2010.

[13] A. Atabati, M. Alizadeh, H. Schuh, and L.-C. Tsai, "Ionospheric scintillation prediction on S4 and ROTI parameters using artificial neural network and genetic algorithm," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2092.

[14] X. Zhao et al., "The prediction of day-to-day occurrence of low latitude ionospheric strong scintillation using gradient boosting algorithm," *Space Weather*, vol. 19, no. 12, 2021, Art. no. e2021SW002884.

[15] N. Linty, A. Farasin, A. Favenza, and F. Dovis, "Detection of GNSS ionospheric scintillations based on machine learning decision tree," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 1, pp. 303–317, Feb. 2019.

[16] A. Dey, M. Rahman, D. V. Ratnam, and N. Sharma, "Automatic detection of GNSS ionospheric scintillation based on extreme gradient boosting technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2021, Art. no. 8014605.

[17] B. C. Vani, M. H. Shimabukuro, and J. F. G. Monico, "Visual exploration and analysis of ionospheric scintillation monitoring data: The ISMR query tool," *Comput. Geosciences*, vol. 104, pp. 125–134, 2017.

[18] L. Spogli et al., "Assessing the GNSS scintillation climate over Brazil under increasing solar activity," *J. Atmospheric Sol.-Terrestrial Phys.*, vol. 105, pp. 199–206, 2013.

[19] P. O. de Camargo, J. F. G. Monico, and L. D. D. Ferreira, "Application of ionospheric corrections in the equatorial region for L1 GPS users," *Earth, Planets Space*, vol. 52, no. 11, pp. 1083–1089, 2000.

[20] G. Seemala and C. Valladares, "Statistics of total electron content depletions observed over the south American continent for the year 2008," *Radio Sci.*, vol. 46, no. 5, pp. 1–14, 2011.

[21] R. de Jesus et al., "Morphological features of ionospheric scintillations during high solar activity using GPS observations over the south American sector," *J. Geophysical Res.: Space Phys.*, vol. 125, no. 3, pp. 1–20, 2020.

[22] B. A. Riwanto, T. Tikka, A. Kestilä, and J. Praks, "Particle swarm optimization with rotation axis fitting for magnetometer calibration," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 2, pp. 1009–1022, Apr. 2017.

[23] C. Denardini et al., "The embrace magnetometer network for south America: Network description and its qualification," *Radio Sci.*, vol. 53, no. 3, pp. 288–302, 2018.

[24] J. King and N. Papitashvili, "Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data," *J. Geophysical Res.: Space Phys.*, vol. 110, no. A2, pp. 1–8, 2005.

[25] A. L. C. d. Souza and P. d. O. Camargo, "Comparison of GNSS indices, ionosondes and all-sky imagers in monitoring the ionosphere in Brazil during quiet and disturbed days," *Boletim de Ciências Geodésicas*, vol. 25, no. SPE, 2019, Art. no. e2019s005.

[26] B. W. Reinisch, "New techniques in ground-based ionospheric sounding and studies," *Radio Sci.*, vol. 21, no. 3, pp. 331–341, 1986.

[27] I. A. Galkin, B. W. Reinisch, X. Huang, and G. M. Khmyrov, "Confidence score of ARTIST-5 autoscaling," Ionosonde Network Advisory Group, USRI, Ghent, Belgium, *INAG Tech. Memorandum*, 2013. Accessed: Feb. 24, 2022. [Online]. Available: https://www.ursi.org/files/CommissionWebsites/INAG/web-73/confidence_score.pdf

[28] M. A. Abdu, "Day-to-day and short-term variabilities in the equatorial plasma bubble/spread f irregularity seeding and development," *Prog. Earth Planet. Sci.*, vol. 6, no. 1, pp. 1–22, 2019.

[29] J. Sousasantos, A. de Oliveira Moraes, J. H. Sobral, M. T. Muella, E. R. de Paula, and R. S. Paolini, "Climatology of the scintillation onset over southern Brazil," *Annales Geophysicae*, vol. 36, no. 2, pp. 565–576, 2018.

[30] V. V. Demyanov, M. A. Sergeeva, and A. S. Yasyukevich, "GNSS high-rate data and the efficiency of ionospheric scintillation indices," in *Satellites Missions and Technologies for Geosciences*. Norderstedt, Germany: Books on Demand, 2019, pp. 1–19.

[31] K. M. Pesyna, T. E. Humphreys, R. W. Heath, T. D. Novlan, and J. C. Zhang, "Exploiting antenna motion for faster initialization of centimeter-accurate GNSS positioning with low-cost antennas," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 4, pp. 1597–1613, Aug. 2017.

[32] A. de O. Moraes et al., "The variability of low-latitude ionospheric amplitude and phase scintillation detected by a triple-frequency GPS receiver," *Radio Sci.*, vol. 52, no. 4, pp. 439–460, 2017.

[33] G. Sivavaraprasad, V. Deepika, D. SreenivasaRao, M. R. Kumar, and M. Sridhar, "Performance evaluation of neural network TEC forecasting models over equatorial low-latitude indian GNSS station," *Geodesy Geodynamics*, vol. 11, no. 3, pp. 192–201, 2020.

[34] G. J. Mendis, J. Wei-Kocsis, and A. Madanayake, "Deep learning based radio-signal identification with hardware design," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2516–2531, Oct. 2019.

[35] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[36] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.

[37] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019.

[38] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of hyperparameters of machine learning algorithms," *J. Mach. Learn. Res.*, vol. 20, no. 53, pp. 1–32, 2019.

[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2623–2631.

[40] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[41] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[42] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLoS One*, vol. 9, no. 2, 2014, Art. no. e87357.

[43] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, 2004, Art. no. 066138.

[44] A. Borisov, V. Eruhimov, and E. Tuv, "Tree-based ensembles with dynamic soft feature selection," in *Feature Extraction*, Berlin, Germany: Springer, 2006, pp. 359–374.

[45] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of Database Systems*, L. Liu and M. T. Ozsu, Eds. Boston, MA, USA: Springer, 2009, pp. 532–538.

[46] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017.

[47] S. Arlot et al., "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40–79, 2010.

[48] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Inf. Sci.*, vol. 191, pp. 192–213, 2012.

[49] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, 2003.

[50] B. Trawiski, M. Smetek, Z. Telec, and T. Lasota, “Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms,” *Int. J. Appl. Math. Comput. Sci.*, vol. 22, no. 4, pp. 867–881, 2012.

[51] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[52] F. Chollet et al., “Keras,” 2015. [Online]. Available: https://keras.io

[53] I. Kutiev et al., “Solar activity impact on the Earth’s upper atmosphere,” *J. Space Weather Space Climate*, vol. 3, 2013, Art. no. A06.

[54] I. Cherniak, I. Zakharenkova, and A. Krankowski, “Approaches for modeling ionosphere irregularities based on the TEC rate index,” *Earth, Planets Space*, vol. 66, no. 1, 2014, Art. no. 165.

[55] Z. Mošna, P. Šauli, and O. Santolık, “Analysis of critical frequencies in the ionosphere,” in *Proc. WDS’Proc. Contributed Papers*, 2008, pp. 172–177.

[56] R. Kaka, E. Nymphas, S. Eniafe, and A. Alabi, “Variations of the critical frequency of the F2 layer, foF2 in west Africa using ionosonde stations at Ouagadougou and Dakar,” *Res. J. Appl. Sci.*, vol. 7, pp. 474–480, 2012.

[57] D. Amorim, A. Pimenta, J. Bittencourt, and P. Fagundes, “Long-term study of medium-scale traveling ionospheric disturbances using oi 630 nm all-sky imaging and ionosonde over Brazilian low latitudes,” *J. Geophysical Res.: Space Phys.*, vol. 116, no. A6, pp. 1–7, 2011.

[58] P. Abadi, Y. Otsuka, and T. Tsugawa, “Effects of pre-reversal enhancement of E × B drift on the latitudinal extension of plasma bubble in southeast Asia,” *Earth, Planets Space*, vol. 67, no. 1, pp. 1–7, 2015.

[59] M. T. Muella et al., “Climatology and modeling of ionospheric scintillations and irregularity zonal drifts at the equatorial anomaly crest region,” *Annales Geophysicae*, vol. 35, no. 6, pp. 1201–1218, 2017.

[60] S. Priyadarshi, Q.-H. Zhang, Y. Ma, Z. Xing, Z.-J. Hu, and G. Li, “The behaviors of ionospheric scintillations around different types of nightside auroral boundaries seen at the Chinese Yellow River station, Svalbard,” *Front. Astron. Space Sci.*, vol. 5, 2018, Art. no. 26.

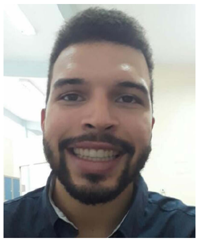
[61] E. R. de Paula et al., “Ionospheric irregularity behavior during the Sep. 6–10, 2017 magnetic storm over Brazilian equatorial–low latitudes,” *Earth, Planets Space*, vol. 71, no. 1, 2019, Art. no. 42.

[62] K. Iwai, D. Shiota, M. Tokumaru, K. Fujiki, M. Den, and Y. Kubo, “Development of a coronal mass ejection arrival time forecasting system using interplanetary scintillation observations,” *Earth, Planets Space*, vol. 71, no. 1, 2019, Art. no. 39.

[63] P. Nogueira et al., “Modeling the equatorial and low-latitude ionospheric response to an intense x-class solar flare,” *J. Geophysical Res.: Space Phys.*, vol. 120, no. 4, pp. 3021–3032, 2015.

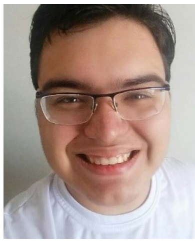
[64] J. Rogers and S. Gunn, “Identifying feature relevance using a random forest,” in *Proc. Int. Stat. Optim. Perspectives Workshop “Subspace, Latent Struct. Feature Selection”*, 2005, pp. 173–184.

[65] J. Jiang, R. Wang, M. Wang, K. Gao, D. D. Nguyen, and G.-W. Wei, “Boosting tree-assisted multitask deep learning for small scientific datasets,” *J. Chem. Inf. Model.*, vol. 60, no. 3, pp. 1235–1244, 2020.



Otávio Carvalho was born in Brazil. He received the bachelor’s degree in science and technology and the master’s degree in aerospace engineering from the Federal University of Maranhão, São Luís, MA, Brazil, from 2019 and 2021, respectively.

He research interest includes artificial intelligence and ionospheric science.

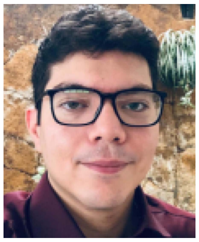


Pedro Augusto Araujo da Silva de Almeida Nava Alves was born in Brazil. He received the bachelor’s degree in computer science from the Federal University of Maranhão, São Luís, MA, Brazil.

His research interests include computer science, with an emphasis on data science, working mainly on the following topic: design of experiments.

Ricardo Yvan de La Cruz Cueva was born in Perú. He received the bachelor’s degree in physics from the Federico Villarreal National University, San Miguel, Perú, in 2006, the master’s degree in physics from the Federal University of Rio Grande do Norte, Natal, MA, Brazil, in 2008, and the Ph.D. degree in space geophysics from the National Institute for Space Research, São José dos Campos, SP, Brazil, in 2013.

He was with the Institute for Scientific Research, Boston College, Chestnut Hill, MA, USA, in 2010. He held a Postdoctoral position with Mackenzie University, São José dos Campos, SP, Brazil, in 2015. His research interests include space geophysics and aeronomy areas.



Alex Oliveira Barradas Filho was born in Brazil. He received the graduate degree in information systems from the University Center of Maranhão, São Luís, MA, Brazil, in 2006, and the master’s degree in electricity engineering and the Ph.D. degree in electricity engineering from the Federal University of Maranhão, São Luís, MA, Brazil, in 2009 and 2015.

His research interests include computer science, with an emphasis on computer systems and artificial intelligence.