Available online at www.sciencedirect.com
ScienceDirect
Advances in Space Research 73 (2024) 3833–3844
ADVANCES IN SPACE RESEARCH
(a COSPAR publication)
www.elsevier.com/locate/asr

Machine learning approach for detection of plasma depletions from TEC

Chandan Kapil, Gopi K. Seemala *

Indian Institute of Geomagnetism, Navi Mumbai 410218, India

Received 6 November 2022; received in revised form 19 April 2023; accepted 25 April 2023
Available online 2 May 2023

Abstract

Ionospheric delay is of concern for *trans*-ionospheric radio communication, especially for the navigation systems relying on these satellite signals. Most of the ionospheric delays are estimated to a degree of first-order using dual frequency global navigation satellite system (GNSS) receivers except during irregularities and equatorial plasma bubbles. The plasma bubbles are observed as a decrease or reduction in total electron content (TEC) as a result of large-scale irregularities that are generated at the equatorial ionosphere. These plasma bubbles can be detected from TEC values visually or by using mathematical algorithms. The mathematical algorithms may have limitations based on assumptions made for the current dataset. Therefore, various machine learning (ML) techniques were tried by training them with selected TEC depletions that are verified for accuracy. From this study, the Random Forest Method (RFM) has performed well compared to other ML methods. The RFM is trained to use for the detection of TEC depletions from the Indian low-latitude region. The training accuracy obtained is 97.6%, with a minimum classification error of 0.023%. The result obtained from the confusion matrix ascertained that the proportion of positively classified cases that are truly positive that is the positive predictive value (PPV) is 96.8%. These statistical results are validated after plotting the observed TEC depletion patch obtained from the ML method. There are cases in which depletions detected by the ML method are appreciable over the mathematical algorithm. The ML technique once trained will not have inherent limitations, as there are no assumptions or threshold values needed to set as required by most of the mathematical algorithms. Thus, the results are encouraging and have scope for further improvement and advancement.

Keywords: TEC depletion detection; Plasma bubble detection; Random Forest method; Machine learning classification techniques

1. Introduction

The ionized plasma of the ionosphere refracts radio waves and affects their propagation. The study of the propagation of radio waves for many researchers is of interest for many decades (Chapman, 1951; Martyn, 1935; Rawer 1993). In general, the day-to-day variation of the ionosphere is well known which affects the propagation of such waves (Rawer 1993). The ionosphere affects satellite-based navigation and communications systems by degrading the signal quality. Most of the parameters that are affected by the ionosphere can be reduced or estimated to first order by using the total electron content, except those due to ionospheric scintillations and irregularities. The equatorial plasma bubble (EPB)/Equatorial spread F (ESF) is one such event, which is observed during post-sunset hours in the F-region of the equatorial and low-latitude ionosphere. This phenomenon leads to the formation of ionospheric irregularities. The formation of these irregularities has been studied for many decades (Basu et al., 1978; McClure et al., 1977; Tsunoda, 1980). The Earth's magnetic field is horizontal at the equator. Thus, the eastward electric field generated by the Pre-Reversal Enhancement (PRE) along with

* Corresponding author.
E-mail address: gopi.seemala@gmail.com (G.K. Seemala).

https://doi.org/10.1016/j.asr.2023.04.042
0273-1177/

the horizontal geomagnetic field rises the plasma irregularities well over 1000 km above in altitude (Huang and Kelley 1996). The rise of plasma irregularities generated in the equatorial ionosphere is dominated by the fountain effect, also known as E $\times$ B drift (Basu and Basu, 1993).

The study of EPB blossomed in ionospheric physics as it is responsible for the degradation of radio wave propagation, which further leads to navigational errors in satellite communication (Seemala and Valladares, 2011). The dynamics of the equatorial ionosphere are dominated by the PRE of vertical drift (Woodman and LaHoz, 1976). During sunset, plasma densities and dynamo electric fields in the E region diminish due to the absence of the Sun. In the F-region a dynamo electric field is generated. Polarization charges, set up by the conductivity gradients at the terminator further increase the eastward electric field for a few hours after sunset. This enhancement of the eastward post-sunset electric field enhances the Rayleigh- Taylor (R-T) instability, while a westward field diminishes it. These structures can grow to become large-scale ionospheric depletions, often called equatorial plasma bubbles (EPB), and are seen as TEC depletions in GPS datasets (Hysell, 2000; Fejer, 1997; Kelley, 1989). These bubbles have been known by different names like 'bite outs', 'depletions', 'plasma holes', and 'ESFs' (Hanson and Sanatani, 1973; McClure et al., 1977). The EPBs/ESFs manifest themselves as diffused echoes in the ionograms, plumes in VHF radar maps, intensity bite-outs in nighttime OI 630 nm airglow intensity, density depletions in the in-situ satellite measurements, and scintillations in VHF/L- band satellite beacon signals (Farley et al., 1970; Hysell and Burcham, 1998; Kelley, 1989; Krishna Moorthy et al., 1989; McClure et al., 1977; Ossakow, 1981; Arons, 1982; Rama et al., 2006; Weber et al., 1978; Woodman and LaHoz, 1976).

The onset of equatorial plasma bubbles has been statistically studied by TEC (total electron content) depletions using GNSS data sets. The TEC of the ionosphere is an important ionospheric parameter that can be used for ionospheric correction in radio systems (Seemala and Valladares, 2011). TEC is a number of electrons along a tube of one-meter squared cross-section. The TEC unit is TECU, 1 TECU = 10^{16} electron /m^2. TEC can be measured by GNSSdata, which is widely used for monitoring the ionosphere (Beach, 2002). TEC depletions are defined as a decrease in the TEC values compared with the surrounding region as observed from GNSS satellite ionospheric pierce point (IPP) footprint trace and are usually observed with apparent durations of 5 min to 40 mins in low latitudinal regions like India (Rama Rao et al., 2006). This decrease in TEC values is followed by the recovery of the values after the depletion. The TEC depletions that we are considering are produced by plasma bubbles drifting across the line of sight between the GPS receiver and the satellite (Qiang Li et al; 2021). Weber et al. (1996) showed a strong correlation between TEC depletions and airglow 630.0 nm depletions, implying that TEC depletions are manifestations of equatorial plasma depletions. The general morphology of TEC depletions and their association with strong levels of VHF scintillations were described by DasGupta et al., (1983) for the pre-midnight period. Yeh et. al., (1979) reported similar correlations in the occurrence of bubbles and scintillations.

To identify plasma bubbles in TEC depletion the threshold for TEC value is generally set to 4TECU. Generally, the TEC depletions greater than 4TECU is due to plasma bubbles according to some previous studies. Such threshold excludes the impact if any due, to other disturbances like Travelling Ionospheric Disturbances (TIDs) (Seemala and Valladares, 2011; Blanch et al., 2018). The TEC depletions have been widely studied using GNSS datasets over the years. The abrupt decrease in the TEC values further followed by apparent recovery as observed in TEC values from the GNSS satellite track is commonly known as TEC depletion(s) at low and equatorial regions. Various methods for detecting plasma bubbles or depletions based on GNSS TEC data have been studied. A few of these are fourth-order polynomial fit (Rahmanmi et al., 2020), moving average (Denardini et al., 2020), band-pass filter technique (Seemala and Valladares, 2011), rolling barrel algorithm method (Freeshah et al; 2020), Savitzky-Golay filter (Osei-Poku et al., 2021; Press et al; 1992), etc. Thus, the efficacy of such methods solely depends upon the assumptions made and thresholds used by the algorithm (Qiang Li et. al., 2021). The diurnal trend or changes in TEC due to disturbed conditions may also be detected as depletions by the mathematical algorithms as mentioned earlier.

With the advancement of Artificial intelligence and the ability to achieve tedious, error-prone tasks, deep machine learning techniques play a crucial role in data-driven space experiments. In past recent years, there has been a resurgence in neural networks and machine learning techniques (Fong et al., 2018). Many researchers have achieved impressive tasks on a variety of challenging problems and analyses using such techniques (Habarulema et al., 2007, 2018; Okoh et al., 2016; Tulsi Ram al., 2018). Machine learning and deep neural networks are the subsets of artificial intelligence, which learn from the input datasets. These ML techniques understand the structure of the data and fit the data into the model. These models are available to the end users to carry out further results by using data (Mcgovern and Wagstaff, 2011). There are three learning methods of ML broadly one is supervised learning which is used in this study, unsupervised learning, and reinforcement learning. The input datasets in supervised learning are labeled, whereas in unsupervised learning the given datasets are unlabelled (Aggarwal, 2018).

Space weather professionals can employ machine learning techniques for real-world applications. Real advances in space weather can be gained using non-traditional approaches that consider nonlinear and complex dynamics. ML has been used in forecasting, now-casting space weather indices (Camporeale et al.,2018), and modeling TEC during geomagnetic storms (Uwamahoro and

Habarulema, 2015) data using complex neural networks. On the other hand, in the presented work, much more sophisticated supervised ML classification techniques such as the Random Forest Method are used to detect the patterns of EPBs manifested in TEC values using GPS/GNSS data. In this study, the machine learning method is used for the detection of ionospheric plasma depletions and the results are discussed and compared with a mathematical algorithm.

1.1. Introduction to Machine learning classification

Machine learning classification techniques have been the most used computational development in the last decade (Rashid, 2016). One of the supervised classification techniques is the Random Forest Method (RFM) also known as Random Forest Ensemble (Herrera et al; 2019). The mechanism of RFM is explained in later sections. RFM algorithms are much more flexible in terms of robustness towards outliers and noise in datasets; thus, the method can predict classes with much more accuracy. The method is simple and easily parallelized. More specifically, ensemble learning algorithms like RFM are well suited for medium to large datasets (Breiman, 2001). Moreover, predictor variables in RFM can be used all at once whereas in logistic and linear regression this cannot be achieved if the number of independent variables is more than observations, that is the number of parameters to be estimated is more than the number of observations (Schonlau and Zou, 2020). In the current study, we have explored the different types of ML techniques for the detection of plasma depletions from TEC data such as K-Nearest Neighbours(K-NN), Decision Tree, Support Vector Machines (SVM), and Random Forest Method. Among these, the Random Forest Method (RFM) classification in ML has resulted in better accuracy in detecting TEC depletions. The RFM algorithm has been successfully applied to reducing high-dimensional and multi-source data in many scientific realms like environmental science, economic recession, etc. (Cavicchioli et al; 2019, Chen and Ishwaran 2012). We aim to explore the efficacy of such an algorithm in the detection of TEC depletions in GPS/GNSS receiver data.

1.2. Random Forest algorithm

As described in the introduction section 1.1, different types of ML algorithms were tried during this study, namely K-NN, SVM, and decision trees. These methods are used basically for classification, and hence can be used for pattern recognition (Rashid, 2016). Based on the accuracy and true positive cases obtained through the respective confusion matrixes of the techniques mentioned, we found that the RFM performs better than the other ML methods for depletion detection. To make a distinction among the methods we must consider certain points which include overfitting problems, features correlation, dimensions of featured space, etc. In RFM, the present estimates can work with missing data. Missing values are substituted by the variable appearing the most in a particular node. It can automatically balance data sets when a class is more infrequent than other classes in the data. The method also handles variables fast, making it suitable for complicated tasks.

Random Forest methods (RFM) have gained considerable importance in machine learning techniques because of their efficient discriminative classification. It is the most used supervised machine learning technique. The RFMis an assembly of classifiers whose core is based upon a decision tree algorithm. RFM uses bagging ensemble learning in which the predictions are based on the combined results of the individual tree. The main goal under the RFM is to segment original data into subsets based upon binary tests to get purer subsets that can be classified as one class.

Let X be the entire dataset of the training phase as shown in step 1 of Fig. 1. In ensemble learning ‘n’ decision trees (Tn) are created from data X. The Tn decision trees represented in step 2 of Fig. 1 are obtained by shuffling rows of the data X, which means they are taken randomly with replacement from the entire dataset X. Thus, the data size of each decision tree is the same. Further, a decision tree classifier from the tree T1 is constructed by splitting the T1 data set into subsets (where X1 is shown as a subset in Fig. 1 as an example) by performing a binary test (C). A binary test is a simple condition such as X1 < C, where C represents any characteristic, and X1 is the input vector at the new node now. The node from where the decision tree starts is called a root node which contains all the datasets. The node where the chain ends is known as the “leaf” node. So, the decision tree works as a flow chart from the root node to the leaf node. There are Tn number of decision trees as represented in step 2 of Fig. 1, each of which is split into nodes, and branches as shown for one decision tree X1 in step 3. The process will continue till each of the purer class forms is obtained that means till the pure class of zeroes or ones is obtained. The purity at each node is calculated through entropy to produce a purer class at the next node (Ali et al; 2012, Zhukov et al., 2018). Finally, at the end of the process, the final prediction is made by voting from all the decision trees (Berhane et al., 2018).

2. Data and methodology

In this study, the data is collected from GPS receivers, at Mumbai and Hyderabad stations. These receivers are ionospheric scintillation monitors (ISM) deployed under the GAGAN network (Acharya et al., 2007) in collaboration with the Airport Authority of India (AAI). The GAGAN network has about 25 stations and was established to study ionospheric characteristics in the Indian region for the implementation of a Space Based Augmentation System (SBAS) for GNSS positioning (Kapil et al., 2022). Mumbai and Hyderabad stations from the GAGAN network are chosen in this present study. Both are low-latitude stations that lie near the inner edge or within the equatorial ioniza-

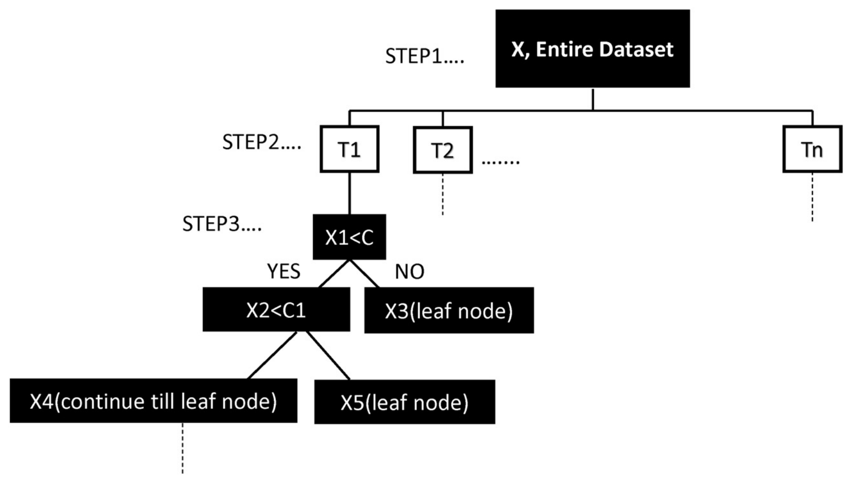

Fig. 1. The chart explains the working of RFM, and step 3 explains the binary test in which X1 is compared with binary characteristics, depending upon which the X1 dataset is segregated in a purer(X3) or impure(X2) data set (Schonlau and Zou, 2020).

tion anomaly (EIA) region. The March 2014 data is chosen for the mentioned stations. The year 2014 was a high solar active year of the solar cycle 24 and in the Indian region, plasma depletions and scintillation occurrence are observed more in the equinoctial months (Chakraborty et al., 2014). The TEC data sampling is 30 sec that is recorded by the ISM receiver. The slant TEC values are converted to vertical TEC using a thin shell mapping function with an ionospheric pierce point altitude of 350 km (Coco et al., 1991; Wilson and Mannucci, 1993; Rama Rao et al., 2006). Hereafter, the vertical TEC is referred to as TEC in the rest of the text.

2.1. Depletion detection

As described in the introduction, the TEC depletions are generally detected/identified using algorithms like detrending, bandpass filtering methods, etc. In the present work, we have used two different techniques for the detection of TEC depletion. The first method, a bandpass filter-based algorithm depletion detection, uses filtered TEC values to identify depletion structures that are beyond set threshold values as described in the later part of this section. The second method, a machine learning-based algorithm that was trained by using carefully selected samples of the former method as inputs, to have a reliable and unbiased method for the detection of depletions that does not depend on the assumed/manually selected threshold values.

In a low latitudinal region like Mumbai, TEC depletions are observed for a duration of 5 to 40 min (Rama Rao et al., 2006). To avoid false detection due to rapid fluctuation in TEC from the background TEC we used two band-pass filters (BPF), the first filter with a cut-off periodicity from 3 to 20 min in the frequency range from 5.5 mHz to 0.83 mHz for narrower depletions and the second BPF with a cutoff frequency from 10 to 50 min that is the frequency range from 1.6 mHz to 0.33 mHz for detection of broader depletions as shown in Fig. 3. The elevation mask angle of 30 degrees is used to avoid noise and multipath effects (Seemala and Valladares, 2011). As shown in Fig. 2, the TEC depletion has occurred during post-sunset can be verified along with the presence of scintillation represented by the S4 index in the lower panel. The elevation angle for this satellite is represented in the black line in the top panel. As EPBs are mainly post-sunset to post-night events thus a time constraint of 12.30 UT to 23.00 UT hours is used, which corresponds to the Indian standard time of 18:00 to 04:30 h (or local time). After this, outliers in the TEC data are removed using a median absolute deviation method (Leys et al., 2013) with a window length of 1 h, as similarly used by Kapil et al., 2022.

Fig. 3 represents the detection of TEC depletions using a digital bandpass for the GPS PRN 2 as observed from Hyderabad on the day of 03 March 2014. The top panel shows the TEC values in a blue solid line and the elevation angle is represented in a black dashed line. And the vertical green and pink lines indicate the onset and end of the respective depletions, as detected by the algorithm using the band pass filtered values. The second panel shows the values of the S4 index representing scintillation activity, and we can observe the scintillation activity simultaneously with the depletions as indicated in the top panel. The third panel shows the values from the first bandpass filter in a range of 3 to 20 min for the detection of narrow or small-size depletions as discussed above, and the fourth panel shows the values from the second filter for the wider depletions in a range of 10 to 50 min. In the third panel, filtered TEC values are in blue, and the red line is the threshold value of filtered TEC chosen for the detection of the depletion. Once, the filtered value decreases or goes below the threshold value, the depletion minima or the point of minimum density is identified and marked as a red asterisk in the third panel. The second step involves identifying the edges (start & ending) of the depletion by finding the values where a change of slope happens and the value is above zero on both sides of this minimum

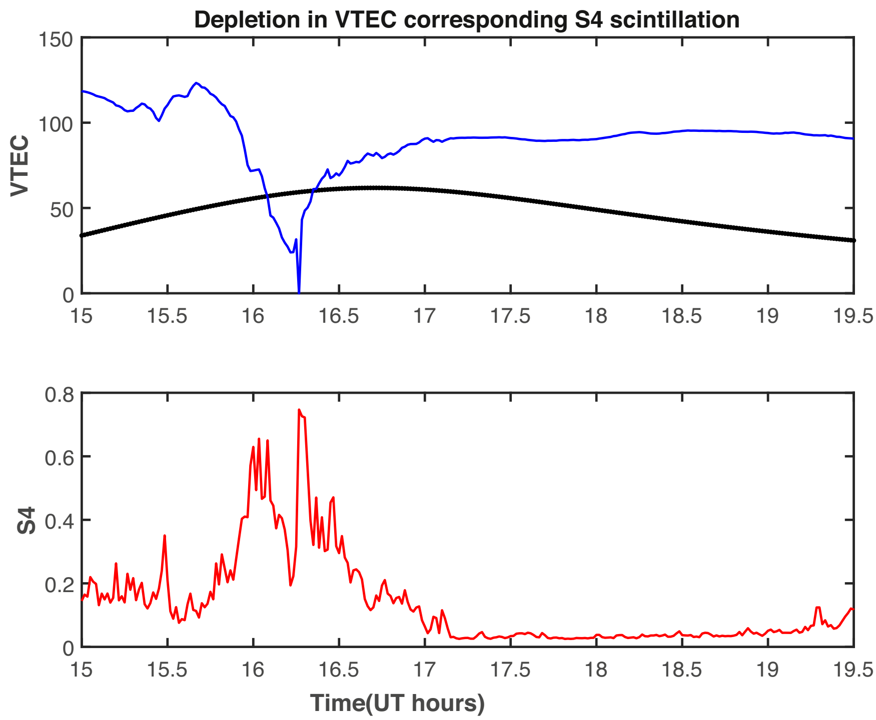

Fig. 2. In this figure, the above panel describes TEC on the y-axis versus time on the x-axis in UT hours in a blue solid line whereas the black line describes the elevation angle. The lower panel describes the S4 index on the y-axis and time (UT hours) on the x-axis in a red solid line.

point. The leading edge and trailing/ending edge of the depletions are marked by green and pink asterisks respectively as shown in the third panel.

The fourth panel shows the filtered TEC values of the second bandpass filter. As discussed above, the same procedure is followed to identify the broader depletions but with a different chosen threshold value. In this current study, we have chosen −1.5 TEC units for the first filter and −3.0 TEC units for the second filter as threshold values for the detection of TEC depletion. The detected depletions are also validated with local time limits i.e., from post-sunset till post-midnight when the equatorial depletions usually occur (Tsunoda et al., 2018). This mathematical technique based on filtering and set thresholds will be called a Filter based Technique (FT) for depletion detection in this paper.

As known, there are limitations with these mathematical techniques as they depend on the type of de-trending method used, filter bandwidths, detection thresholds used, etc. Hence, we have resorted to machine learning by giving the best-detected samples as input using the traditional band pass algorithm. The ML algorithm takes the training from the detected depletion samples which are manually selected. The core behavior of the ML technique is that the algorithm once trained can detect depletions in the cases of the testing phase for the current period of study (2014) without needing any manual inputs like threshold levels and filter bandwidth like in this case of a mathematical algorithm. These manual inputs would limit the detection capability to a particular set of anticipated depletions of depth and duration. To prepare a training data set from the traditional band pass method, any false detection of TEC depletion is checked visually for each sample.

The depletion detection (mathematical) algorithm needs manual intervention when it must be applied to the data at a different location or solar activity. As the algorithm requires tuning of the threshold values, and the bandpass filter cut-off periodicities that are specific to the ambient ionization densities (TECs) and depletion durations expected. The TEC values and depletions durations may change with the location, i.e., whether it is closer to the magnetic equator or at the equatorial ionization anomaly region. Effectively, the threshold used in this algorithm must be tuned for the chosen station and also partially depending on the solar activity owing to the level of ambient ionization produced. To avoid this manual tuning of threshold values, a ML algorithm can be trained with the data and detected samples.

To train the RFM algorithm as described in section 1.2, the selected data with detected depletions using the mathematical algorithm are given as inputs in the training phase. The detected depletions are classified as a series of 1 s, and the rest of the time when there is no depletion is classified as a series of 0 s as shown in Fig. 4. The input training data is prepared as described in section 3. We used the classification learner package provided in MATLAB for training the RFM algorithm (Paluszek and Thomas, 2017). The differ-

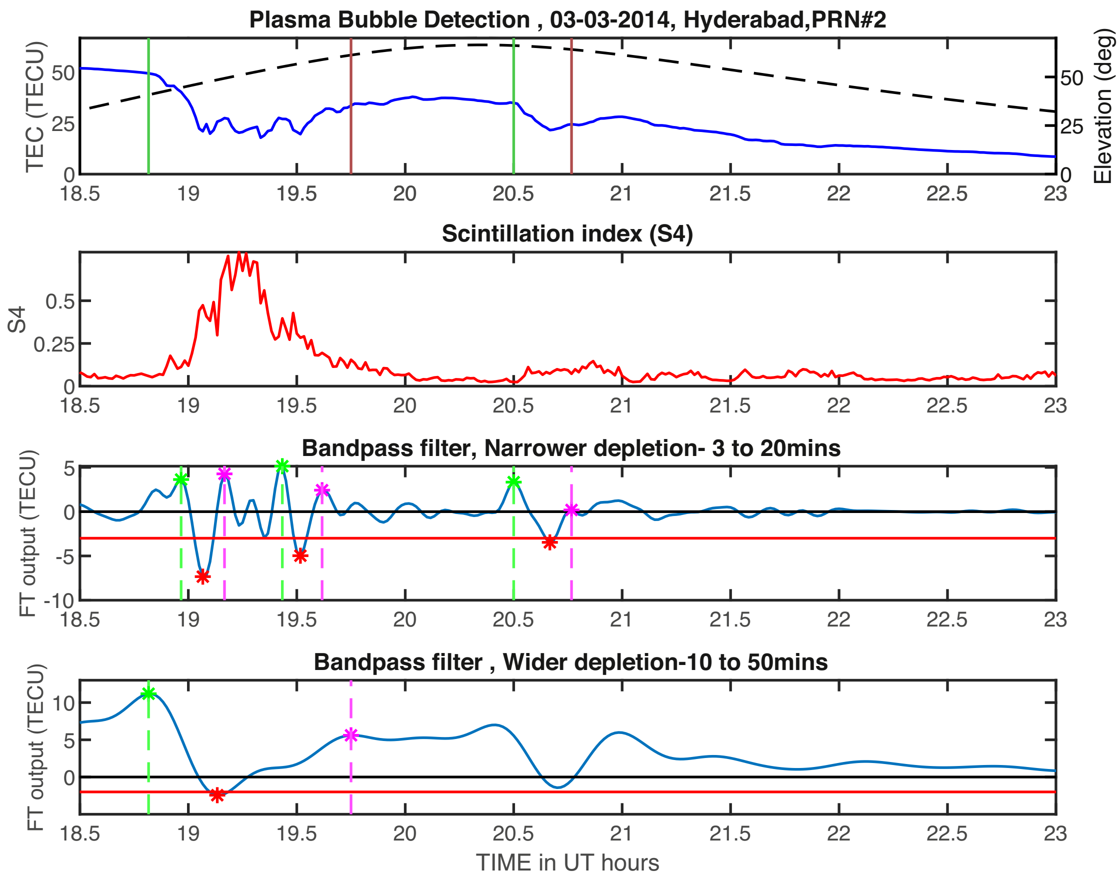

Fig. 3. In the top panel, the blue line represents the TEC and the black dashed line is the elevation angle, green and brown lines represent the start and end time of depletions. In the second panel, the red line represents the corresponding scintillation index (S4). The third and fourth panels show filtered TEC values after the narrow band pass filter (3 to 20 mins) and wide band pass filter(10 to 50 mins) respectively. The green and magenta dashed lines represent the start and end time of depletions respectively, the red asterisks represent the maximum depth of the depletions. Also in the third and fourth panels, the red solid line represents the threshold cut-offs used for the detection of depletions.

ent types of parameters in addition to the TEC that were used for training the RFM are discussed in the following section.

3. Training dataset for ML technique

To prepare the training data set for ML, 66.5% of the data is chosen from the data as mentioned in section 2. The remaining 33.5% dataset is reserved for testing and validation of the trained ML model, which is independent of the dataset that is given for training. The training dataset is pre-processed using the FT method (as described in section 2.1) to detect depletions to use the outcome for the training of the ML model. From these detected TEC depletion samples, the best samples are visually chosen so that accurate training can be imparted to ML. A set of binary logic values are generated to represent the presence of detected depletion for the duration by setting the values to 1′s, and the rest of the data where no depletion is detected are indicated by values of 0′s as shown in panels A of Fig. 4 along with corresponding FT outputs in panels B and C respectively. These labeled binary logic values will aid in the training of machine learning for TEC depletion detection. Thus, for training, each of the detected samples is plotted as shown in Fig. 4 for visual inspection to aid in selection. The red solid line in Fig. 4 represents the classification of TEC depletion by zeroes and ones. The presence of depletion is indicated by the value of 1′s and the 0′s represents the absence of depletion. Thus, the square/rectangular portion between the leading and trailing red lines indicates a depletion. As can be seen in Fig. 4a, there is no depletion and it is represented with 0′s. In Fig. 4b, two depletions are detected and are represented by the 1′s as shown by a red line.

In addition to TEC values, the features that are selected are time, PRN of satellite, day of the year, latitude & longitude of IPP, F10.7, and Dst Index as the input for ML. Thus, the training dataset includes 8 available features to classify the TEC depletion for training along with labeled binary logic as described above. To get a fair estimation of model performance we use a 5-fold cross-validation technique (Baturynska et al., 2021) while training with the dataset. There are 126 detected depletions and 40 non-depletion cases that were given in the training dataset for ML.

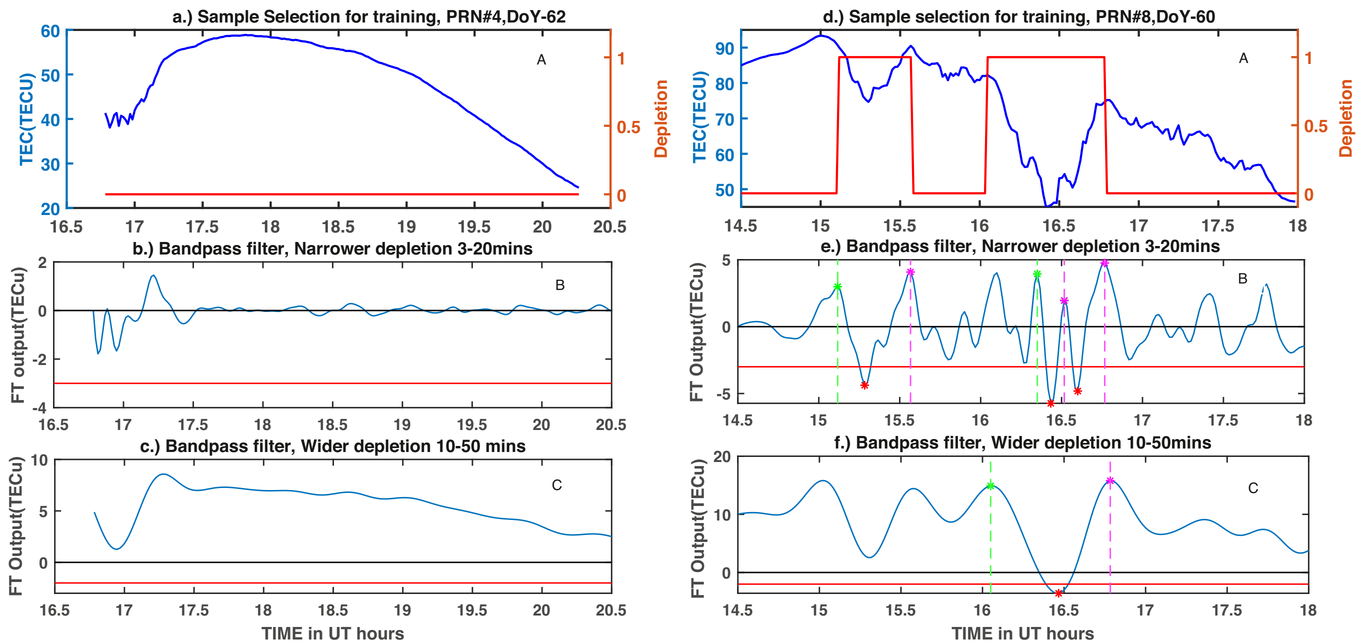

Fig. 4. The depletion detection samples from the FT method are marked by red solid lines in panel A of both figures. The left panel shows the TEC values (blue curve) and no depletion is indicated by the flat red solid line. The right panel shows the identified depletions marked by 1′s (or high) and the absence of depletion region by 0′s (or low). Panels B and C in both figures show the filter output for the narrow band and wide band respectively as described in section 2.1.

4. Results

4.1. Performance of training model

To understand the performance of such models given in MATLAB and Python packages we must understand the terms on behalf of which the model's accuracy is evaluated. To check the performance of the trained model we need to evaluate it, to implement it on the testing dataset. For that, we consider Area Under the Curve of Receiver Characteristic Operator (AUC-ROC). The AUC-ROC curve is the evaluation metric for binary classification problems. It is defined as the probability curve that plots the true positive rates (TPR) against the false positive rates (FPR) at different threshold values. The curve helps us to visualize how better the technique has performed in the training phase. As we have made our problem a binary classification problem after assigning zeroes and ones to the non-depletion and depletion times series respectively. Further TPR is also known as sensitivity, defined as $\frac{TP}{TP+FN}$, here True Positives, False Positives, False Negatives, and True Negatives are represented as TP, FP, FN, and TN respectively. FPR is known as specificity which is defined $\frac{TN}{TN+FP}$. So, the AUC-ROC curve is the curve between sensitivity and 1- specificity. The values of the AUC-ROC curve can vary between 0 and 1, tending toward 1 means the training models are distinguishing between all the positives and negatives correctly (Bradley, 1997). To better understand the categories, we must investigate the confusion matrix, which is our second parameter to evaluate the model. A confusion matrix is a table that counts between the predicted and actual values. Fig. 5 depicts the diagram to understand the confusion matrix in a better way. The labels TN, FP, FN, and TP are the same as described above. Every row of the matrix links to the actual class and every column to the predicted class in Fig. 5. The total counts of correct and incorrect classifications are entered into the table as shown in Fig. 5. The sum of correct predictions for a class goes into the predicted column and expected row for that class value, similarly for incorrect predictions (Duntsch and Gediga et al., 2019).

In the present study, the confusion matrix obtained from the model is presented in Fig. 6. In Fig. 6, the first and second column describes false (0) and true (1) cases of the predicted class respectively whereas the first and sec-

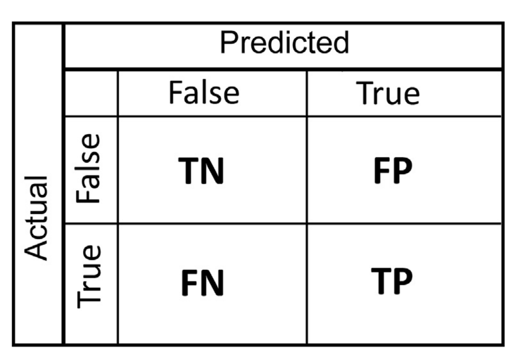

Fig. 5. The schedule as confusion matrix predicts the number of total TN, FP, FN, and TP.

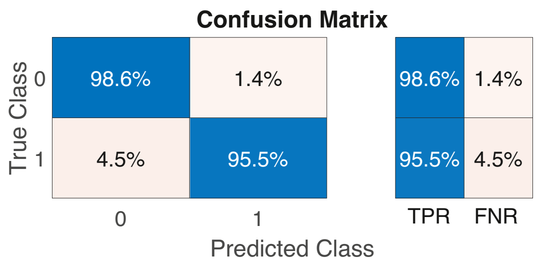

Fig. 6. Confusion matrix obtained by RFM which corresponds to Fig. 5.

ond rows describe true (1) and false (0) cases of the actual class. It shows between the two classes of zeroes and ones, the account of TN cases are 98.6%, FP cases are 1.4%, FN cases are 4.5% and TP cases are 95.5%. In Fig. 6, TPR describes true positive rates and FNR describes false negative rates. Thus, we can see that the model training data set can predict the maximum of actual true positive cases of depletions and true negatives of no depletions. To ensure this we also check the AUC-ROC curve as explained above shown in Fig. 7.

We obtained that ROC has a good separability among the binaries at the threshold represented by a red dot that is at (0.04, 0.99). The area under the curve obtained at the threshold mentioned for the present study is 1 which tells that the classifier perfectly distinguishes between the zeroes and ones (Matjaz and Zoran, 2013).

The minimum classification error plot for the trained model is obtained and is shown in Fig. 8. A predefined hyperparameter optimization can be automated within the ML algorithm. The main advantage to optimize these hyperparameters by selecting an optimizable ensemble is that the classification method will tune its parameters to reduce the classification error and retain the optimal setting for internal parameters like the number of splits, distribution, ensemble method, etc. The hyper-parameters optimized are mentioned in the legend of Fig. 8. Fig. 8 illustrates the minimum classification error for RFM. The convergence curve that is the dark blue line refers to the observed minimum classification error computed so far by the optimized model. While light blue is the convergence curve when examining all the hyper-parameters values. At the last, a global minimum error is obtained with the ensemble method named Gentle Boost. The number of learners is 484 at a learning rate of 0.040916. The maximum number of splits is 1751. The red square shows the minimum error obtained at the optimized hyperparameter named as best point hyper-parameter, the value of the minimum error is precisely marked with a yellow circle shown in Fig. 8.

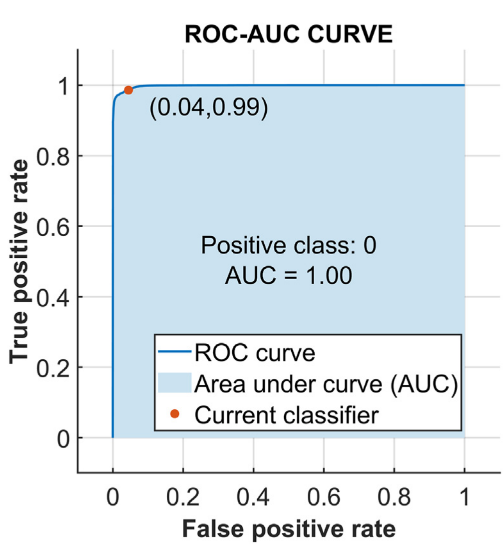

Fig. 7. AUC-ROC generated after the RFM training phase.

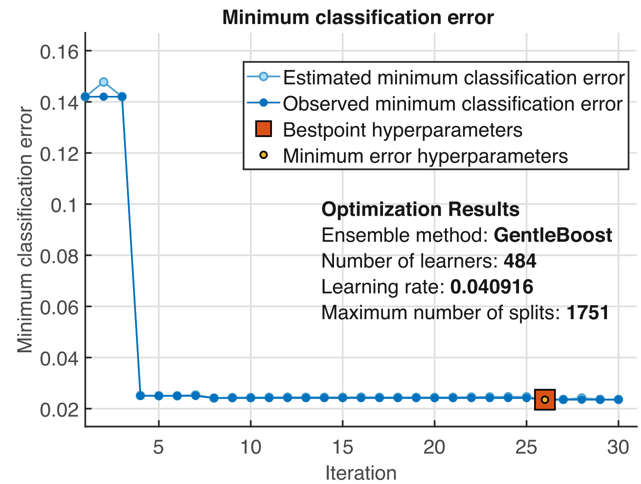

Fig. 8. Minimum classification error obtained from the RFM after the training phase.

These two measures i.e., confusion matrix and ROC, described above will give us the performance of the classifier; thus from the above Figs. 6 & 7, we have obtained a better classification result from the training set. It took around 8 h approximately for the training phase to complete.

In this present case, rather than the accurate prediction of these binary classified depletion indicators, it is more important to check the accurate depletion detection in terms of its start and end times from the given dataset. Therefore, the ML-detected depletion is compared with FT-detected depletions with different samples in the following section.

4.2. Test results and discussion

To check the efficacy of the ML classification method, we need to test the trained model for its accuracy in the detection of TEC depletion which is indicated by the binary values. Though the ML method has shown good performance on the trained dataset as shown in Figs. 6 & 7, it is important to check whether the entire depletion is detected correctly instead of each instance or epoch that is identified accurately. Therefore, the TEC depletions detected by ML must be compared with FT-detected depletion samples as well as visually examined.

As discussed in section 3, only the best output samples from FT are used for training, so any inaccurate depletion detections were not used for training and hence may not propagate to ML output or ML output will not produce artifacts based on these inaccurate samples. Now, after training the ML method, we have chosen the remaining independent dataset (not used for ML training) as described in section 3, for the validation of the trained ML. The trained ML model is used to detect depletions from this dataset for validation. In most cases, it is observed that the ML technique works well, especially if there is single depletion to be detected. Fig. 9 shows the detection results obtained from a few cases using both FT and ML models. Fig. 9 is composed of six sub-figures (a,b,c,d,e, & f). There are five panels in each figure, and the topmost panel in each Fig. 9(a,b,c,d,e, & f) represents the TEC values in a solid blue line. Panels (ii & iii) show the detrended TEC values that are used by the FT method for the detection of depletions in each of the figures. The panels (iv & v) represent the depletions detected by the FT method in the red solid line as explained above (in section 2.1) and the bottom panels represent the depletions detected by the ML method in the solid purple line.

From Fig. 9, it can be observed that in a few cases, the accurate detection of TEC depletion using ML is improved compared to the FT method, for example, in Fig. 9 (a, b, c & d). In Fig. 9(a), the duration of the depletions detected by ML is better than that of FT as can be visually seen. Similarly, in Fig. 9(b & f), the duration of the depletion detected by ML is improved but the second depletion was not detected by ML. In Fig. 9(c), FT has detected single depletion with multiple structures in it as two depletions, but ML has detected it as single depletion.

In Fig. 9(d), the FT method has detected 3 depletions, as per the threshold described in section 2.1, but the ML method could detect only single depletion probably as it had encountered a negative spike in the dataset as can be seen in the top panel. Similarly, in Fig. 9(e), the ML method could detect a depletion but the ML method has also produced artifact spikes at about 16 h and 17 h, probably because of the spike in the dataset (observed in the top panel). In Fig. 9(d & e), the TEC values have some artifact fluctuation either due to noise or data gap due to loss of lock/cycle slip correction, in this case, ML detected depletion but also gave some erroneous values or spikes. Otherwise, in both cases Fig. 9 (d & e), the duration of depletion is improved in ML compared to the FT method apart from spikes as can be observed from TEC in top panels.

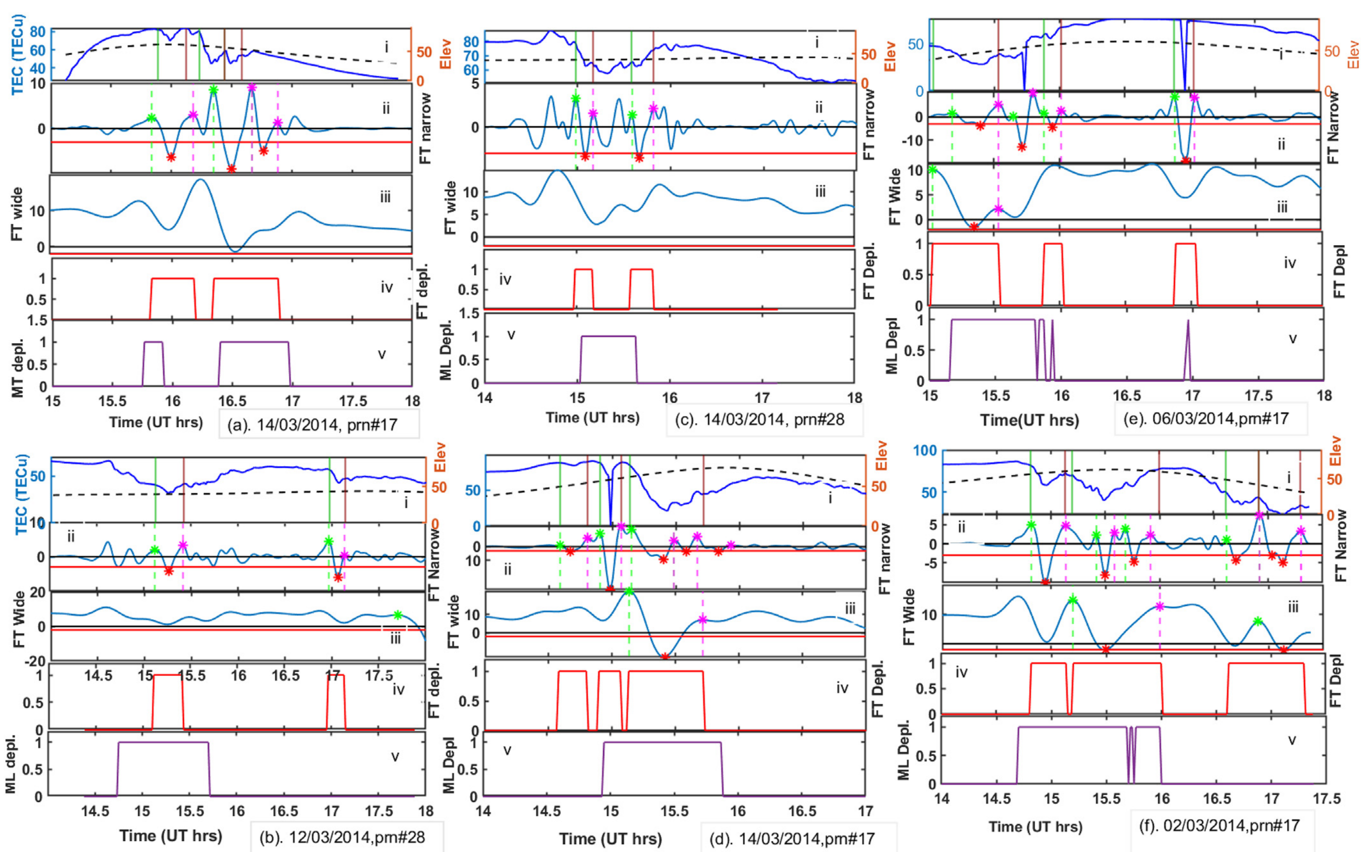

Fig. 9. Comparison of the detection of plasma bubbles using FT as well as ML. The top panel (i) of each sub-figures shows TEC in the blue line. Panels (ii & iii) show narrow-band and wide-band filter outputs respectively that are used by FT. The green and magenta vertical dashed line describes the start and end of each depletion respectively in panels (ii & iii). Panels (iv & v) show depletions detected by the FT method (in red line) and detected by the ML method (in purple line) respectively.

Therefore, from observation of multiple cases, the ML has detected the TEC depletion regions with some improvement in time duration in comparison to the FT method as observed visually from the TEC values. Here, we have taken the bandpass algorithm as a reference to check the efficacy of the ML technique on the other hand each case is also checked visually with original TEC data in the first panel as shown in Fig. 9. In very few cases like multiple artifacts or discontinuity in TEC values, the current trained ML has detected the depletions but also some isolated 1′s or spikes are seen.

The major advantage of the ML technique is that it does not require manual assumptions/inputs like fixing threshold values for the detection or changing the filter bandwidths to accommodate different depletion durations. As we have seen from Fig. 9 (ii & iii), ML has accurately detected the depletions with multiple sub-structures in it where the FT accuracy is less as visually observed. The ML method can be improved by training it with more complex cases and more data points, as discussed by many ML researchers (Sarker et al., 2021, Najafabadi et al., 2015). The error in the detection of depletions is 15.2 percent in the current test dataset, i.e., 61 were detected by ML out of 72 depletions which are visually examined. Here, this error in the ML method is defined as the inaccurate detection of the entire depletion, as shown in Fig. 9 and discussed above. Thus, the ML method has an advantage over the FT method, i.e., once trained it is simple and easy to be used for future analysis. However, the ML has a limitation, if the training datasets are not large enough, it may not be trained properly and so the model will not be accurate. By increasing the size of training dataset, the model becomes more general by virtue of being trained on more examples and could decrease the error in the identification of TEC depletions. The results may vary depending upon the significant difference in the solar activity period chosen.

5. Summary

Real-world complex problems are comprised of non-linear systems mostly. In such systems, the output is not proportional to the inputs given. To solve such systems or to better understand them ML is a new technical approach. The variation in TEC depending upon ionospheric condition is one such non-linear phenomenon. The data sets of Earth’s ionospheric space are typically imbalanced especially near EIA regions, due to the availability of many days of quiet conditions and a few hours of storms. This poses a serious problem in any mathematical algorithm that tries to find patterns in the data. It is also problematic for defining meaningful metrics that assess the ability of a model to predict interesting events. When ML techniques are applied with the proper amount of data sets, we can obtain better sensitivity and separability.

In the current study, the detection of TEC depletions is considered, which are day-to-day varying events and are highly dependent on various electrodynamic environmental factors. In this study, we tried to evaluate ML efficacy to detect TEC depletions using data from low latitude GPS stations Hyderabad and Mumbai from March 2014, the results are quite significant. We have used RFM technique to detect the TEC depletions which are compared to FT as shown in Fig. 9. The training accuracy obtained in RFM is 97.6%, with a minimum classification error of 0.023 percent, and truly positive cases that is the positive predictive value (PPV) is 96.8 percent. We have observed an error of about 15.21 percent in the identification of TEC depletions using the ML with the given training dataset. Thus, the ML using RFM detects the TEC depletions quite well with a scope of future improvements as discussed in the result’s section. With ML methods, if large datasets are not given for training, it may not be trained perfectly. Hence with ionospheric datasets, the results may vary depending upon the significant differences in ambient TEC owing to changes in solar activity. Thus, adding more datapoints increases the training accuracy. Future works in this field can be focused on the development of feature selection and fuzzy logic for performance enhancements.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the Indian Institute of Geomagnetism (IIG), Department of Science and Technology, India. The authors acknowledge the high-performance computing facility at IIG for the extended services and support. The authors thankfully acknowledge Space Application Center (SAC), ISRO Ahmedabad, for providing GPS data from the GAGAN network repository. The GAGAN Data is available through Acharya and Majumdar (2019). One of the authors (C.K.) is grateful to IIG for the fellowship awarded towards her Ph.D.

References

Acharya, R., Majumdar, S., 2019. Statistical relation of scintillation index S4 with ionospheric irregularity index ROTI over Indian equatorial region. Adv. Space Res. 64 (5), 1019–1033. https://doi.org/10.1016/j.asr.2019.05.018.

Acharya, R., Nagori, N., Jain, N., Sunda, S., Regar, S., Sivaraman, M.R., Bandopadhya, K., 2007. Ionospheric studies for the implementation of GAGAN. Indian Journal for Radio Space Physics 36, 394–404.

Aggarwal, C.C., 2018. Recurrent Neural Networks. In: Neural Networks and Deep Learning. Springer International Publishing, Switzerland, Vol.10 (978), pp. 271-313. ISBN 978-3-319-94463-0. doi.org/10.1007/978-3-319-94463-0

Arons, J., 1982. Global morphology of ionospheric scintillations. In: Proceedings of the IEEE. Vol. 70 (4), pp. 360 – 378. doi.org/10.1109/PROC.1982.12314.

Basu, S.a., Basu, S.u., Aarons, J., McClure, J.P., Cousins, M.D., 1978. On the coexistence of kilometer- and meter-scale irregularities in the nighttime equatorial F region. J. Geophys. Res. 83 (A9), 4219–4226. https://doi.org/10.1029/JA083iA09p04219.

Basu, S.a., Basu, S.u., Tatarski, V.I., Ishimaru, A., Zavorotny, V.U., 1993. Ionospheric structures and scintillation spectra. In: Wave Propagation in Random Media (Scintillation). The International Society for Optical Engineering, Bellingham, WA, USA, pp. 139–153.

Baturynska, I., Martinsen, K., 2021. Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. J. Intell Manuf 32, 179–200. https://doi.org/10.1007/s10845-020-01567-0.

Beach, T.L., 2002. Global positioning system studies of ionospheric irregularities: A technical review. COSPAR Colloq. Ser. 12, 249–254.

Berhane, T.M., Lane, C.R., Wu, Q., Autrey, B.C., Anenkhonov, O.A., Chepinoga, V.V., Liu, H., 2018. Decision-Tree, Rule-Based, and Random Forest Classification of High-Resolution Multispectral Imagery for Wetland Mapping and Inventory. Remote Sens. 10 (4), 580. https://doi.org/10.3390/rs10040580.

Blanch, E., Altadill , D., Miguel, J.J., Camps, A., Barbosa, J., González-Casado, G., Riba, J., Sanz, J., Vazquez, G., Orús-Pérez, R., 2018. Improved characterization and modeling of equatorial plasma depletions. Journal of Space Weather and Space Climate, 8, A38. dx.doi.org/10.1051/swsc/2018026.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Breiman, L., 2001. Random Forests Machine Learning. Nature Springer, Kluwer Academic Publishers, Netherlands 45, 5–32.

Camporeale, E., Wing, S., Johnson, J. R., 2018. Machine Learning Techniques for Space Weather. Elsevier, ISBN 9780128117880. doi.org/10.1016/B978-0-12-811788-0.09994-7.

Cavicchioli, M., Papana, A., Dagiasis, A.P., Pistoresi, B., 2019. A Random Forests Approach to Assess Determinants of Central Bank Independence. J. Mod. Appl. Stat. Methods 172, 3–26.

Chakraborty, S., Chatterjee, S., Dasgupta, A., 2014. SBAS operation in the Indian sector under simultaneous scintillation on different GNSS links. J. Atmos. Sol. Terr. Phys. 114, 50–57. https://doi.org/10.1016/j.jastp.2014.04.005.

Chen, X., Ishwaran, H., 2012. Random forests for genomic data analysis. Genomics, Vol. 99(6), pp. 323-9, 10.1016/j.ygeno.2012.04.003.

Chapman, S., 1951. Some Phenomena of the Upper Atmosphere. Proc. Phys. Soc. B 64, 833. https://doi.org/10.1088/0370-1301/64/10/301.

Coco, D.S., Coker, C., Dahlke, S.R., Clynch, J.R., 1991. Variability of GPS satellite differential group delay biases. IEEE Transaction on Aerospace and Electronic Systems 27 (6), 931–938. https://doi.org/10.1109/7.104264.

DasGupta, A., Basu, S.a., Aarons, J., Klobuchar, J.A., Basu, S.u., Bushby, A., 1983. VHF amplitude scintillations and associated electron content depletions as observed at Arequipa, Peru. J. Atmos. Terr. Phys. 45 (1), 15–26. https://doi.org/10.1016/S0021-9169(83)80003-8.

Denardini, C. M., Picanço, G. A. S., Barbosa Neto, P. F., Nogueira, P. A. B., Carmo, C. S., Resende, L. C. A., 2020. Ionospheric scale index map based on TEC data for space weather studies and applications. Space Weather, 18, e2019SW002328. doi.org/10.1029/2019SW002328.

Düntsch, I., Gediga, G., 2019. Confusion matrix and rough set data analysis. J. Phys.: Conf. Ser. Vol. 1229, 012055, doi.org/10.1088/1742-6596/1229/1/012055.

Farley, D.T., Balsey, B.B., Woodman, R.F., 1970. Equatorial spread F: Implications of VHF radar observations. J. Geophys. Res. Space Phys. 75, 7199–7216. https://doi.org/10.1029/JA075i034p07199.

Fejer, B.G., 1997. The electrodynamics of the low latitude ionosphere: recent results and future challenges. J. Atmos. Sol. Terr. Phys. 59 (13), 1456–1482. https://doi.org/10.1016/S1364-6826(96)00149-6.

Fong, R.C., Scheirer, W.J., Cox, D.D., 2018. Using human brain activity to guide machine learning. Nature Scientific Rep 8 (1), 2045–2322. https://doi.org/10.1038/s41598-018-23618-6.

Freeshah, M., Zhang, X., Jun Chen, J., Zhao, Z., Osama, N., Sadek, M., Twumasi, N., 2020. Detecting Ionospheric TEC Disturbances by Three Methods of Detrending through Dense CORS During A Strong Thunderstorm. Ann. Geophys. 63 (6), GD667. https://doi.org/10.4401/ag-8372.

Habarulema, J., McKinnell, L.A., Cilliers, P., 2007. Prediction of GPS total electron content using neural network over South Africa. J. Atmos. Sol. Terr. Phys. 69 (15), 1842–1850. https://doi.org/10.1016/j.jastp.2007.09.002.

Habarulema, J.B., McKinnell, L.A., Opperman, B.D.L., 2018. A recurrent neural network approach to quantitatively studying solar wind effects on TEC derived from GPS; preliminary results. Ann. Geophys. 27 (5), 2111–2125. https://doi.org/10.5194/angeo-27-2111-2009.

Hanson, W.B., Sanatani, S., 1973. Large Ni gradients below the equatorial F peak. Journal Geophysical Research, Space Physics 78 (7), 1167–1173. https://doi.org/10.1029/JA078i007p01167.

Herrera, V.M., Khoshgoftaar, T.M., Villanustre, F., 2019. Random Forest implementation and optimization for Big Data analytics on LexisNexis's high-performance computing cluster platform. Journal of Big Data 6 (1), 68. https://doi.org/10.1186/s40537-019-0232-1.

Huang, C.S., Kelley, M.C., 1996. Nonlinear evolution of equatorial spread F: On the role of plasma instabilities and spatial resonance associated with gravity wave seeding. J. Geophys. Res. Space Phys. 101 (A1), 283–292. https://doi.org/10.1029/95JA02211.

Hysell, D.L., 2000. An overview and synthesis of plasma irregularities in equatorial spread F. J. Atmos. Sol. Terr. Phys. 62 (12), 1037–1056.

Hysell, D.L., Burcham, J., 1998. JULIA radar studies of equatorial spread F. J. Geophys. Res. 103 (A12), 29155–29167. https://doi.org/10.1029/98JA02655.

Kapil, C., Seemala, G.K., Shetti, D.J., Acharya, R., 2022. Reckoning ionospheric scintillation S4 from ROTI over Indian region. Adv. Space Res. 69 (2), 915–925. https://doi.org/10.1016/j.asr.2021.10.026, ISSN 0273–1177,.

Kelley, M.C., 1989. The Earth's ionosphere, Plasma Physics and Electrodynamics. Academic Press, San Diego, CA.

Krishna Moorthy, K., Prabha, R., Nair Krishna Murthy, B.V., 1989. Multiwavelength solar radiometer network and features of aerosol spectral optical depth at Trivandrum. Indian Journal of Radio Space Physics 18, 194–201.

Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 49, 764–766. https://doi.org/10.1016/j.jesp.2013.03.013.

Majnik, M., Bosnic, Z., 2013. ROC analysis of classifiers in machine learning: A survey. Intell. Data Anal. 17, 531–558. https://doi.org/10.3233/IDA-130592.

Martyn, D.F., 1935. The propagation of medium radio waves in the ionosphere. Proc. Phys. Soc. 47, 323. https://doi.org/10.1088/0959-5309/47/2/311.

McClure, J.P., Hanson, W.B., Hoffman, J.H., 1977. Plasma bubbles and irregularities in the equatorial ionosphere. J. Geophys. Res. 82 (19), 2650–2656. https://doi.org/10.1029/JA082i019p02650.

Mcgovern, A., Wagstaff, K., 2011. Machine learning in space: Extending our reach. Mach. Learn. 84, 335–340. https://doi.org/10.1007/s10994-011-5249-4.

Najafabadi, M., Villanustre, F., Khoshgoftaar, T., Seliya, N., Wald, R., Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. Journal of Big Data 2 (1). https://doi.org/10.1186/s40537-014-0007-7.

Osei-Poku, L., Tang, L., Chen, W., Mingli, C., 2021. Evaluating Total Electron Content (TEC) Detrending Techniques in Determining Ionospheric Disturbances during Lightning Events in A Low Latitude Region. Remote Sens. 13 (23), 4753. https://doi.org/10.3390/rs13234753.

Ossakow, S.L., 1981. Spread F Theories-A review. J. Atmos. Terr. Phys. 43 (5–6), 437–452. https://doi.org/10.1016/0021-9169(81)90107-0.

Paluszek, M., Thomas, S., 2017. An Overview of Machine Learning. In: MATLAB Machine Learning. Apress, Berkeley, CA. doi.org/10.1007/978-1-4842-2250-8_1.

Press, W.H., Teukolisky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The art of scientific computing. Cambridge University Press, Cambridge, U.K..

Qiang Li, Q., Zhu, Y., Wang, Z., Fang, K., 2021. A method for automatic detection and characterization of plasma bubbles using GPS and BDS data. Chin. J. Aeronaut. 34 (5), 195–204.

Rahmani, Y., Alizadeh, M.M., Schuh, H., Wickert, J., Tsai, L.C., 2020. Probing vertical coupling effects of thunderstorms on lower ionosphere using GNSS data. Adv. Space Res. 66, 1967–1976.

Rama Rao P. V. S., S. Seemala, G. K., Niranjan K., Prasad, D.S.V.V.D., 2006. Temporal and spatial variation in TEC using simultaneous measurements from the Indian GPS network of receivers during the low solar activity period of 2004–2005. Ann. Geophys., Vol. 24, pp. 3279-3292, 10.5194/angeo-24-3279-2006.

Rama Rao, P.V.S., Niranjan, K., Prasad, D.S.V.V.D., Gopi Krishna, S., Uma, G., 2006. On the validity of the ionospheric pierce point (IPP) altitude of 350km in the Indian equatorial and low-latitude sector. Ann. Geophys. 24, 2159–2168. https://doi.org/10.5194/angeo-24-2159-2006.

Rashid, T., 2016. Make your own Neural network. CreateSpace Independent Publishing Platform, North Charleston, SC, United States.

Rawer, K., 1993. Propagation through the ionosphere. In: Wave Propagation in the Ionosphere. Developments in Electromagnetic Theory and Applications, Vol. 5, pp:139-158, Dordrecht, Springer. doi.org/10.1007/978-94-017-3665-7_10.

Sarker, I.H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. Sn Comput. Sci. 2, 160. https://doi.org/10.1007/s42979-021-00592-x.

Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. Stata J. 20 (1), 3–29. https://doi.org/10.1177/1536867X20909688.

Seemala, G.K., Valladares, C.E., 2011. Statistics of total electron content depletions observed over the South American continent for the year. Radio Sci. 46, RS5019. https://doi.org/10.1029/2011RS004722.

Tsunoda, R.T., 1980. Magnetic-field-aligned characteristics of plasma bubbles in the nighttime equatorial ionosphere. J. Atmos. Terr. Phys 42, 743–752. https://doi.org/10.1016/0021-9169(80)90057-4.

Tsunoda, R.T., Saito, S., Nguyen, T.T., 2018. Post-sunset rise of equatorial F layer—or upwelling growth? Prog. of Earth Planet Sci 5 (1), 22.

Tulasi Ram, S., Sai Gowtam, V., Mitra, A., Reinisch, B., 2018. The improved two-dimensional artificial neural network-based ionospheric model (ANNIM). J. Geophys. Res. Space Phys. 123, 5807–5820.

Uwamahoro, J.C., Habarulema, J.B., 2015. Modelling total electron content during geomagnetic storm conditions using empirical orthogonal functions and neural networks. J. Geophys. Res. Space Physics 120, 11000–11012. https://doi.org/10.1002/2015JA021961.

Weber, E.J., Buchau, J., Eather, R.H., Mende, S.B., 1978. North-south aligned equatorial airglow depletions. J. Geophys. Res. 83, 712–716. https://doi.org/10.1029/JA083iA02p00712.

Wilson, B.D., Mannucci, A.J., 1993. Instrumental-biases in ionospheric measurements derived from GPS data. Proc. of Ion GPS-93, Inst. of Navigation, Washington, D.C..

Woodman, R.F., LaHoz, C., 1976. Radar observations of F region equatorial irregularities. J. Geophys. Res. 81, 5447–5466.

Yakoov Weber, Y., 1996. Corporate Cultural Fit and Performance in Mergers and Acquisitions. Hum. Relat. 49 (9), 1181–1202.

Yeh, K.C., Soicher, H., Liu, C.H., Borelli, E., 1979. Ionospheric bubbles observed by the Faraday rotation method at Natal, Brazil. Geophys. Res. Lett. 6, 473–475.

Zhukov, A., Sidorov, D., Mylnikova, A., Yasyukevich, Y., 2018. Machine learning methodology for ionosphere total electron content nowcasting. International Journal of Artificial Intelligence 16 (1), 144–157.