# Comparative Analysis of Machine Learning Models for Total Electron Content Reconstruction in the Beijing Area

Qiushuo Wang

December 6, 2024

**Abstract**

The ionosphere is a critical part of Earth's atmosphere, directly influencing GPS navigation and radio communications. Total Electron Content (TEC) is a key parameter that quantifies the state of the ionosphere. Accurate TEC prediction has been a focal point of numerous empirical and physical modeling efforts. With advancements in computational techniques, machine learning models have shown great promise in this field. This paper compares the performance of three models—Artificial Neural Network (ANN), Support Vector Regression (SVR), and Random Forest (RF)—in predicting TEC over the Beijing area. The results indicate that while the ANN model demonstrates superior overall accuracy with high correlation and low mean square error, capturing the general diurnal, seasonal, and annual variations of TEC, the RF model excels in reproducing short-term, fine-grained variations during specific conditions, such as daytime summer TEC changes. This comprehensive evaluation highlights the strengths and limitations of each model, offering insights into their suitability for different aspects of ionospheric TEC modeling.

## 1 Introduction

The ionosphere is a portion of Earth's atmosphere, lying about 50 to 1000 kilometers above the Earth's surface. Here, the energy from solar ultraviolet and X-rays ionizes gas molecules, producing a plethora of free electrons and ions. The presence of these free electrons, particularly their quantity and distribution, significantly impacts the propagation of radio waves. They can alter the speed and path of radio signals, influencing the accuracy of radio communications and Global Positioning System (GPS) readings.

Total Electron Content (TEC) is an integral parameter describing the state of the ionosphere. More specifically, if choosing a vertical plane along the signal path from a point on the Earth to a satellite, then TEC is the number of all electrons within this plane. It is critical for the functionality of GPS, radio communication, and our understanding of Earth's atmosphere. TEC provides a method to quantify the number of free electrons in the ionosphere. Typically, GPS signals or other satellite signals from Earth's orbit are used to calculate TEC. This is because as these signals pass through the ionosphere, they undergo refraction and delay, effects that can be used to estimate the number of electrons. By collecting data from multiple GPS receivers and processing it with the appropriate algorithms, global TEC maps can be created.

The unit for TEC is the TECU, where 1 TECU is equivalent to $10^{16}$ electrons. In TECU, "U" stands for "unit". This unit has been widely accepted in depicting the electron content of the ionosphere, especially in contexts such as error correction for GPS signals.

There are many empirical and physical models to predict TEC. The International Reference Ionosphere (IRI) model, which is recognized as the official standard for the Earth's ionosphere by the International Standardization Organization, the International Union of Radio Science, the Committee on Space Research, and the European Cooperation for Space Standardization, is a data based model that describes average ionospheric values of electron density, TEC, electron and ion temperature, and ion composition (Bilitza et al., 2022). With the development of hardware facilities and update of computing power, Machine Learning (ML) has become a powerful technique to build empirical model. Utilizing this technique, Rukundo et al., 2023 reconstructed the local Egypt TEC, which had high accuracy in diurnal variation compared with IRI model. Chen et al., 2019 utilized the DCGAN algorithm to fill missing data in TEC maps. Sorkhabi, 2021 employed deep learning of artificial neural networks (ANN) to estimate TEC for single-frequency (SF) GPS users.

In this work, I applied several popular ML algorithms, including the artificial neural network (ANN), support vector regression (SVR), and random forest (RF), to model local TEC in Beijing area from the year of 2005 to 2017. These models perform well on the metrics with high Pearson correlation coefficients (R) and low mean square error (MSE), and can successfully reconstruct the annual, seasonal and diurnal variations of TEC.

## 2 Data Description

The data used in this study was preprocessed and normalized by linear interpolation and outlier processing by Mengting et al., 2020, which has 15 grid points at latitudes 5°N, 30°N, 40°N, 50°N, 75°N, and longitude 110°E, 115°E, 120°E, including 6 parameters: TEC, Bz, Kp, Dst, F10.7 and AE. The original TEC data comes from the International IGS Ionospheric Analysis Center CODE, the time, and the Bz, Kp and other data comes from NASA's data access and orbit service OMNIWeb.

In this study, I only use data in grid 115°E 40°N, which is also the location of Beijing area. Only Kp and F10.7 are chosen as input of ML models. The Kp index, with a range of 0 (quiet) to 9 (extreme storm), is one of the most commonly used indices to quantify the disturbance in the Earth's magnetic field caused by solar activity. The F10.7 index, which is a measure of the solar radio flux per unit frequency at a wavelength of 10.7 centimeters, is one of the key indices used to characterize the intensity of solar activity and its impact on the Earth's ionosphere and magnetosphere. This combination is determined by multiple experiments, suggesting that adding extra input parameters may lead to overfitting and worse performance on the test set.

Figure 1 illustrates the temporal variations of TEC, F10.7, and the Kp index from 2005 to 2017. TEC has a time resolution of 1 hour, F10.7 has a daily time resolution, and the Kp index has a time resolution of 3 hours. As Figure 1 shows, there is a strong correlation between F10.7 and TEC, while the correlation between Kp index and TEC is relatively weaker. However, due to the sensitivity of both Kp index and TEC to geomagnetic disturbances, the role of Kp index cannot be ignored (Sorkhabi, 2021).

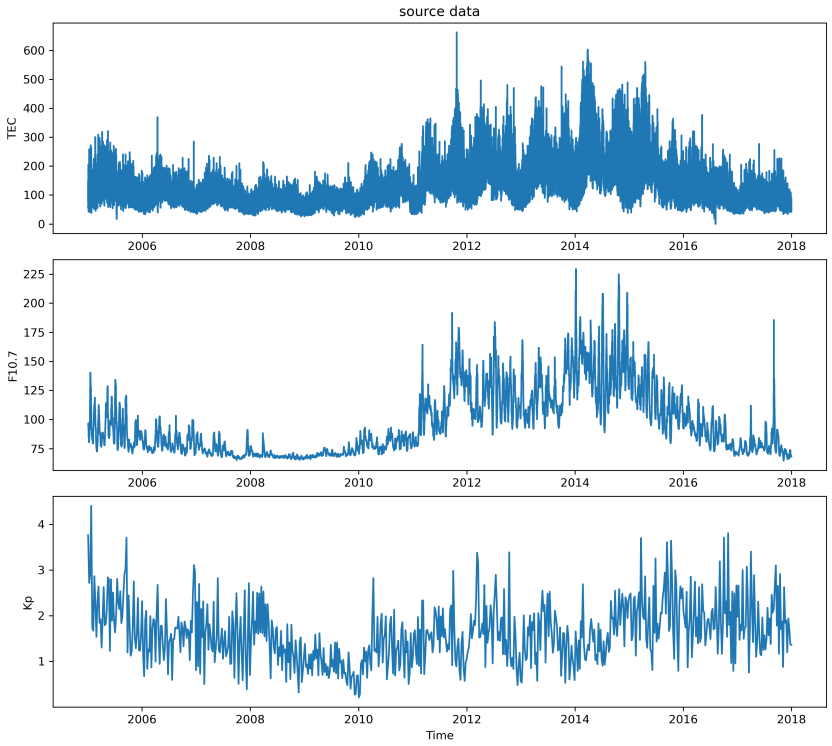


Figure 1: Source Data.

# 3 Model Training

To ensure consistent time resolution, the entire dataset is sampled at an hourly resolution, which means that for F10.7, the same value is repeated every 24 points. Following this sampling approach, a total of 113,208 data samples were obtained.

Besides Dst and F10.7, time parameters

$$HS = \sin \frac{2\pi * h}{24}$$

$$HC = \cos \frac{2\pi * h}{24}$$

$$DS = \sin \frac{2\pi * d}{365.25}$$

$$DC = \cos \frac{2\pi * d}{365.25}$$

are used as input. Here, $h$ represents the hour of the day for local time, $d$ represents the calendar day of year. It is important to add time parameters as input because TEC has significant diurnal and seasonal variations. These time parameters are converted to Sine and Cosine components to maintain continuity (Habarulema et al., 2007).

The data set is divided into a training set (˜70%), a validation set (˜15%) and a test set (˜15%). The validation set is only used for ANN model as a monitor to avoid overfitting. The ANN model adopts 2 hidden layers with 200 neurons per layer, and use the stochastic gradient descent (SGD) optimizer to minimize the loss function (in this case it is the mean squared error (MSE) between predicted and observed values) at each time step to update the

weights and biases. The training process stops either when the MSE of the validation set stops improving for 32 consecutive steps to prevent over-fitting or when the training reach 2000 epochs. These settings and parameters are chosen after multiple experiments and guided by metrics R and MSE.

The SVR model provided is a pipeline combining data preprocessing with StandardScaler and an SVR model using the Radial Basis Function (RBF) kernel. This kernel is particularly effective for modeling nonlinear relationships in data. The SVR is configured with a regularization parameter (C) set to 1.0, which balances the trade-off between achieving a low error on training data and maintaining model generalization. The epsilon parameter (epsilon=0.1) defines a margin of tolerance around the true target values, reducing sensitivity to minor prediction errors. Considering the data size is moderate (˜110,000) and computational efficiency is a concern, SVR is a reliable choice as well.

The RF model provided is an ensemble learning method that builds multiple decision trees and combines their predictions to achieve better accuracy and generalization. This implementation uses 100 trees (n_estimators=100), with no limit on tree depth (max_depth=None) to allow trees to grow until fully split. The model employs the square root of the total features (max_features='sqrt') for each split, helping to reduce overfitting and improve performance on unseen data. RF is robust to overfitting due to their averaging nature and can handle both regression tasks effectively.

# 4 Model Performance

To illustrate the performance of the models, the data set of the year 2014 is used as the test set (not used in the training and validating processes). Figure 2 presents a comparative analysis of the performance of the three regression models in predicting log TEC values in 2014. Each subplot displays a heatmap of sample density, with the x-axis representing the observed log TEC values and the y-axis showing the modeled log TEC values. A red diagonal line is included in each plot to indicate the ideal prediction line where modeled values perfectly match observed values.

The ANN model demonstrates the highest accuracy, with the lowest Mean Squared Error (MSE = 0.021) and the highest correlation coefficient (R = 0.955), indicating a strong agreement between observed and predicted values. The RF model shows moderate performance, with an MSE of 0.048 and R = 0.926, while the SVR model performs the least accurately, with an MSE of 0.060 and R = 0.897. These results suggest that the ANN model is the most effective in capturing the underlying patterns of the data, followed by SVR and RF. Note SVR and RF models significantly underestimate high TEC values, which is evident in the plots, where the predicted values for high TEC are shifted below the diagonal line, particularly for the RF model. The error likely arises because these models struggle to accurately capture the nonlinear relationships and extreme variations in the data.

In the case of SVR, its reliance on support vectors to define the model may cause it to focus on the majority of the data distribution, at the expense of extreme values like those observed during high TEC conditions. Similarly, RF's ensemble averaging nature inherently smooths predictions, leading to reduced sensitivity to high TEC values and a tendency to regress predictions toward the mean.

Such high TEC values are typically associated with intense geomagnetic activity, such as geomagnetic storms, which induce large-scale variations in ionospheric electron density. The

inability of SVR and RF to effectively model these extreme conditions indicates that their performance is less robust during geomagnetically active periods, as compared to the ANN model, which appears to handle these variations more effectively. This suggests that the ANN model's ability to learn complex, nonlinear patterns allows it to better predict TEC across a wider range of conditions, including those influenced by geomagnetic disturbances.

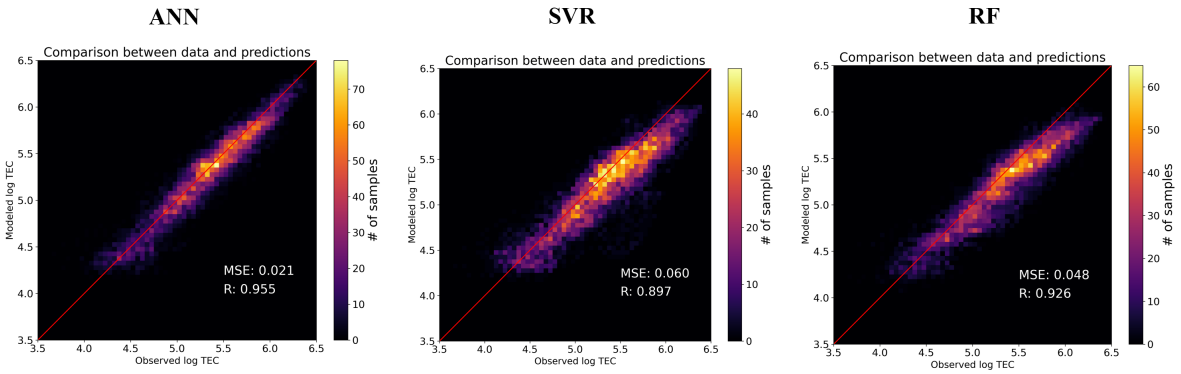


Figure 2: Comparison of model predictions versus observed values for three regression models: Artificial Neural Network (ANN), Support Vector Regression (SVR), and Random Forest (RF). Each subplot shows a heatmap of sample density along with a diagonal red line representing perfect predictions. The performance metrics include Mean Squared Error (MSE) and correlation coefficient (R).

## 4.1 Annual and Seasonal variation

The seasonal and monthly variations in TEC are governed by an interplay of solar activity, solar zenith angles, thermospheric composition, and neutral winds (Kumar and Singh, 2009; Patel et al., 2017; Bagiya et al., 2009). During the equinox, the sun's axis aligns parallel to the equator and subsequently shifts to higher latitudes throughout the summer and winter seasons. Generally, within the mid-latitude region like Beijing, TEC values peak in summer and dip to their lowest in winter. The annual variation in TEC is significantly correlated with the solar activity cycle. For instance, TEC values during periods of high solar activity, such as from 2010 to 2012, substantially exceed those during periods of low solar activity, like from 2005 to 2007.

Figure 3 provides a visualization of the predicted versus actual average TEC values for the entire dataset (February 1, 2005, to December 31, 2017) using these three ML models. All of these models successfully reconstructs the annual and seasonal variations, accurately capturing the low and high values across different seasons, and captures the variation of TEC in both solar high years and solar low years. However, none of the models successfully predict TEC values during extreme geomagnetic conditions, such as the significant geomagnetic storm in early September 2017. This limitation highlights a shared weakness across the models in capturing the highly dynamic and nonlinear behavior of TEC under such extreme conditions. Besides, the RF model demonstrates a clear underestimation of TEC values during the first half of 2014 (test set). This behavior aligns with the observations from Figure 2, where RF consistently underpredicts high TEC values. This underestimation is particularly evident in periods of elevated TEC, reflecting the RF model's tendency to regress toward the mean and its inability to accurately model high TEC peaks.

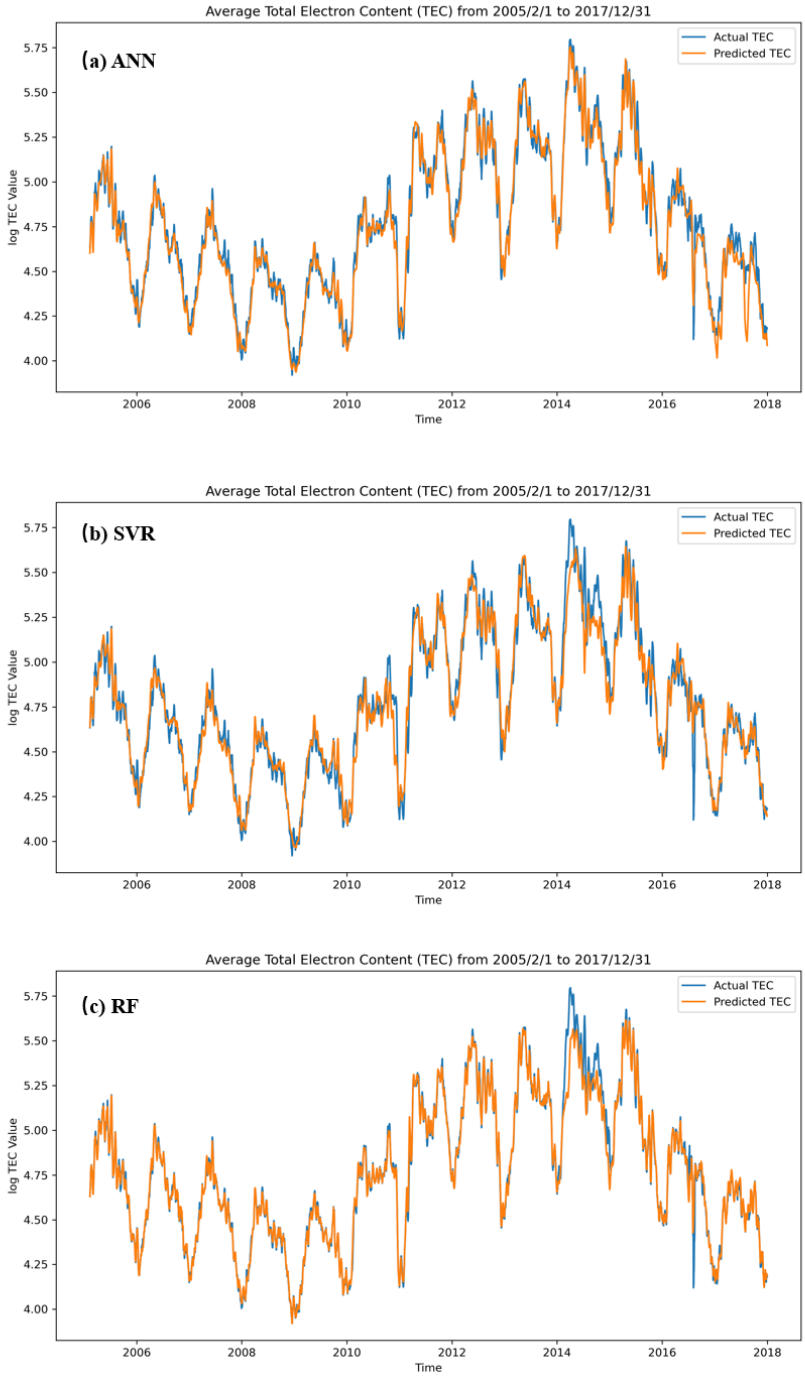


Figure 3: The predicted versus actual average TEC values (in log scale) for the entire dataset (from February 1, 2005, to December 31, 2017) using: (a) Artificial Neural Network (ANN), (b) Support Vector Regression (SVR), and (c) Random Forest (RF). The blue lines represent the actual TEC values, while the orange lines indicate the predicted TEC values.

## 4.2 Diurnal Variation

The diurnal variation in TEC is contingent upon the rate of photoionization and recombination processes, both of which are modulated by solar activity. The rate of photoionization escalates

in the morning, reaching a maximum around noon before gradually diminishing with the setting of the sun. During the night, in the absence of sunlight, the recombination process predominates, causing TEC to decline to its minimum value around midnight (Olwendo et al., 2016; Oryema et al., 2015). However, the TEC value at midnight doesn't vanish entirely due to the elevation of the F2 layer to higher altitudes with minimal recombination rates, coupled with the contribution of plasmaspheric electrons (Li et al., 2018).

Figure 4 provides a detailed comparison of the predicted versus actual TEC values during a week in the spring of 2014 (test set). All three models successfully capture the diurnal patterns of TEC, characterized by peaks during the day and troughs at night, demonstrating their ability to model the general periodic variations in the data. The ANN model exhibits the closest alignment with the actual TEC values, while the SVR model and RF model underestimates the daytime TEC peaks, although it still follows the general temporal structure, consistent with our previous findings. The same results repeat in fall diurnal variation (see Code on the Website).

For a summer week shown in Figure 5, however, RF model captures the detailed and complex variations in daytime TEC during the summer better than the ANN model, although the RF model exhibits weaknesses in overall predictive accuracy. In contrast, the ANN model primarily captures the broader "daytime increase, nighttime decrease" trend, lacking the finer granularity seen in the RF model's predictions. This result suggests that despite ANN's superior overall performance, as evidenced by the evaluation metrics in earlier analyses, it may oversmooth the TEC variations, missing some of the intricate short-term fluctuations during the summer daytime (and winter midnight, see Code on the Website). The RF model, on the other hand, seems better suited for capturing these small-scale, dynamic changes, particularly in the diurnal cycle in summer and winter. This highlights the RF model's strength in reproducing finer details in the data, even though it struggles with general trends and extreme values. In this regard, the performance of the SVR model lies between the two, capturing more detail than ANN but falling short of RF in terms of reproducing the full complexity of daytime variations.

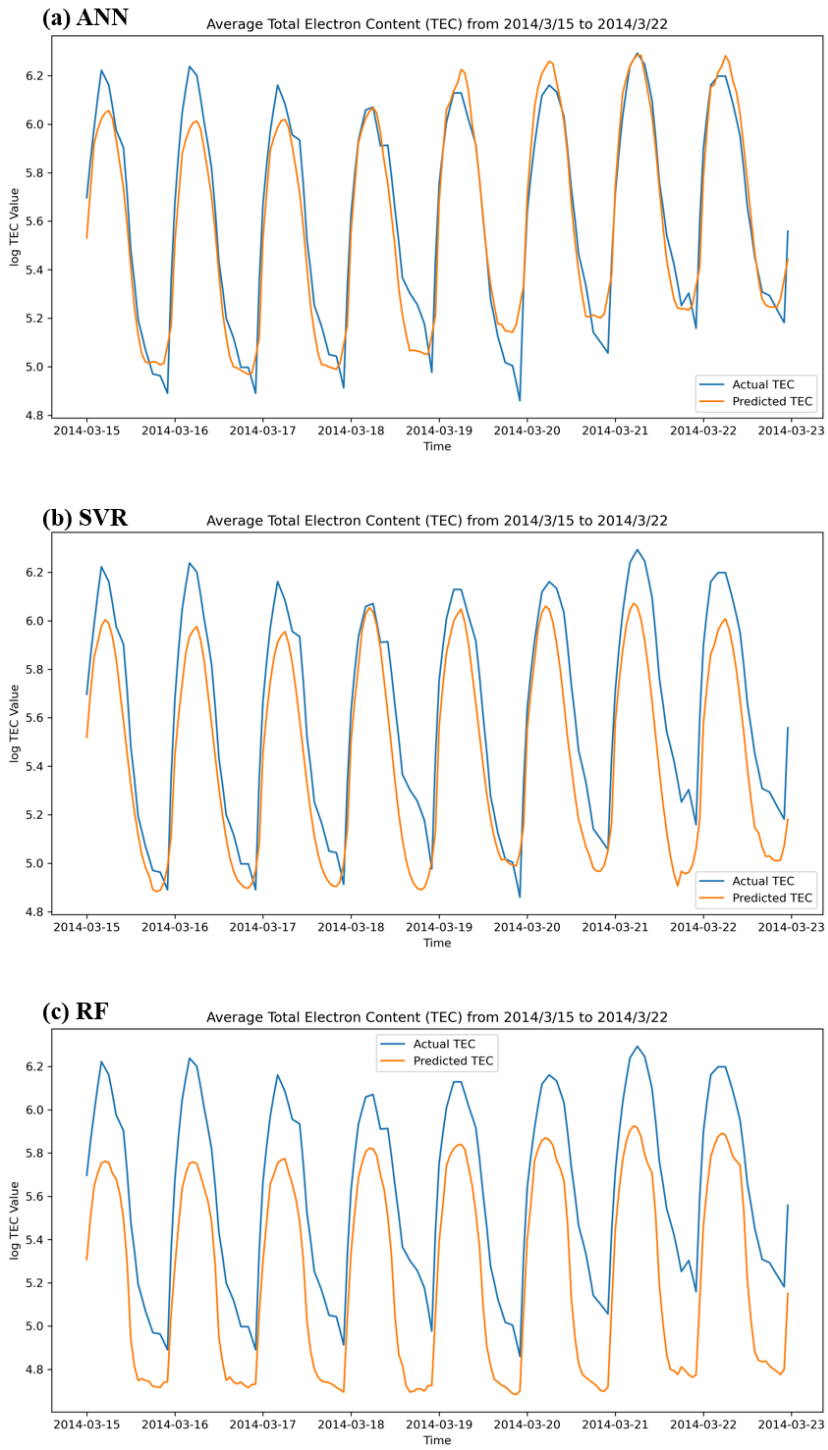


Figure 4: The predicted versus actual average TEC values (in log scale) during a week (March 15–22, 2014) of the spring of 2014 (test set), using: (a) Artificial Neural Network (ANN), (b) Support Vector Regression (SVR), and (c) Random Forest (RF). The blue lines represent the actual TEC values, while the orange lines show the predicted values.

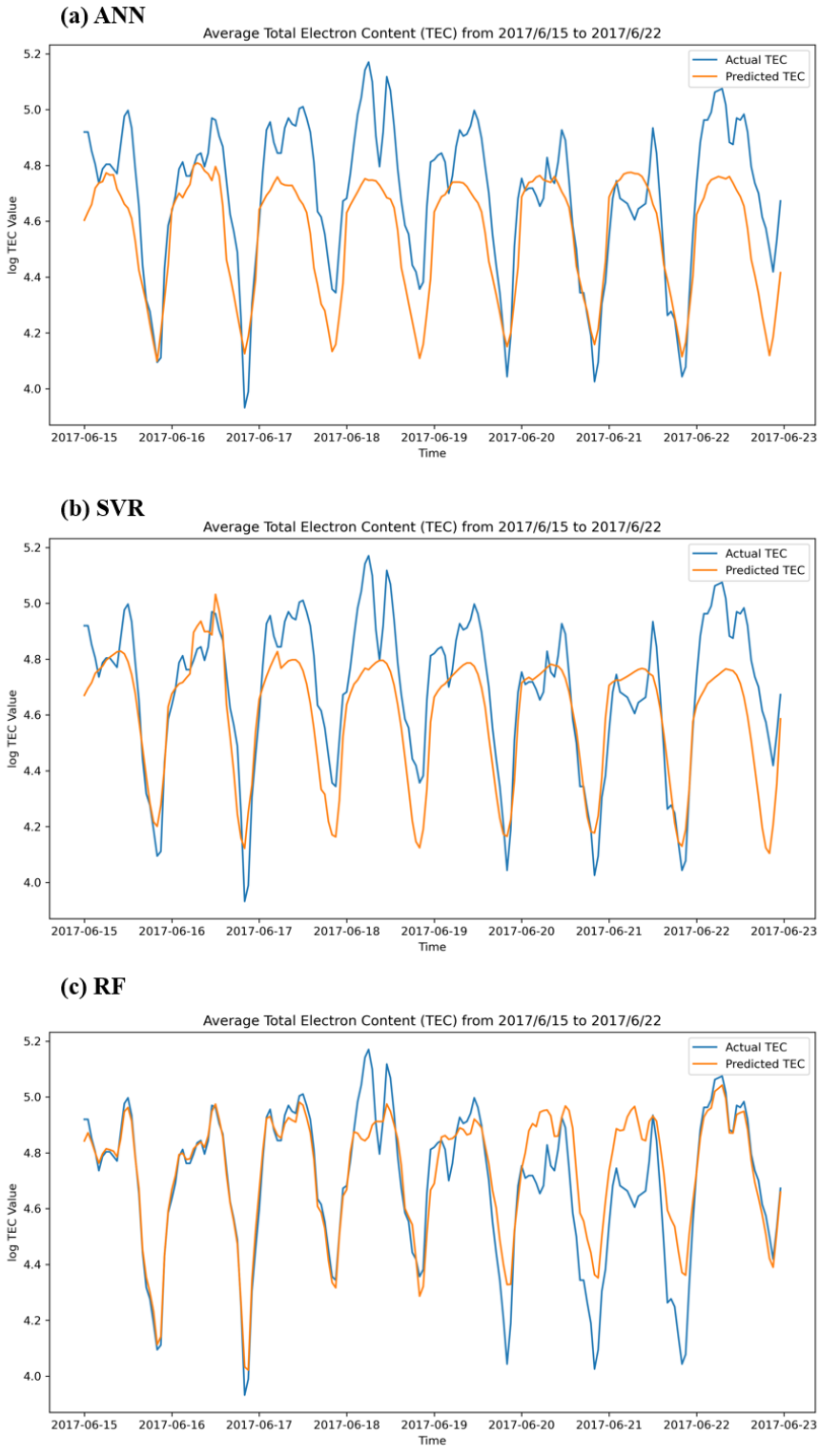


Figure 5: The predicted versus actual average TEC values (in log scale) during a week (June 15–22, 2014) of the summer of 2014 (test set), using: (a) Artificial Neural Network (ANN), (b) Support Vector Regression (SVR), and (c) Random Forest (RF). The blue lines represent the actual TEC values, while the orange lines show the predicted values.

# 5 Discussion and Conclusion

This study presents a detailed comparison of three regression models—Artificial Neural Network (ANN), Support Vector Regression (SVR), and Random Forest (RF)—for predicting Total Electron Content (TEC). The evaluation, conducted over both short-term and long-term periods, highlights the distinct strengths and limitations of each model, providing valuable insights into their suitability for TEC prediction under various conditions.

ANN consistently demonstrates superior performance across most evaluation metrics, achieving the lowest Mean Squared Error (MSE) and highest correlation coefficients. Its ability to capture both the general diurnal TEC variations and high TEC values makes it the most robust model for overall prediction. However, as shown in the summer short-term analysis, ANN tends to oversmooth the predictions, primarily capturing the broad "daytime increase, nighttime decrease" trend while missing finer fluctuations in daytime TEC.

SVR performs moderately well, capturing the general patterns of TEC but struggling with extreme values, particularly during high TEC conditions. Its underestimation of peaks is consistent across both short-term and long-term analyses, suggesting that while it effectively handles mid-range TEC values, it lacks the flexibility to model extreme variations induced by geomagnetic activity or seasonal effects.

RF shows notable strengths in capturing the detailed, dynamic daytime variations in TEC, particularly during the summer season. Despite this, RF exhibits a significant limitation in underestimating high TEC values, as observed during geomagnetically active periods and in the 2014 testing dataset. This tendency to regress toward the mean highlights its limitations in modeling extreme conditions. Nonetheless, its ability to reproduce short-term diurnal fluctuations more accurately than ANN or SVR suggests potential for hybrid modeling approaches.

All models struggled to capture extreme TEC variations during geomagnetic storms, such as the event in September 2017. Applying data augmentation techniques, such as oversampling high TEC events or creating synthetic data, may help balance the dataset. Using weighted loss functions could also prioritize learning from these rare events. Besides, the current ANN architecture oversmooths TEC predictions, missing some finer details, especially during daytime variations. Integrating some detailed ionospheric indices (for example: indices with F2 layer height) may help ANN models capture finer variations of TEC. For RF struggling with capturing high TEC values, transitioning to gradient-boosted decision trees, such as XGBoost, which are more flexible and less prone to averaging effects, might improve performance, particularly in handling high TEC peaks.

Future work may focus on developing hybrid models that combine the robustness of ANN with the fine-grained sensitivity of RF and SVR. For example, using RF or SVR to handle detailed short-term fluctuations, and leveraging ANN for robust overall trend predictions and extreme values. Additionally, exploring advanced architectures such as ensemble learning or physics-informed machine learning could further enhance TEC prediction under extreme conditions, improving operational forecasting and space weather monitoring.

# References


Bilitza, D., Pezzopane, M., Truhlik, V., Altadill, D., Reinisch, B., & Pignalberi, A. (2022). The international reference ionosphere model: A review and description of an ionospheric benchmark. *Reviews of Geophysics*, *60*. https://doi.org/10.1029/2022RG000792

Rukundo, W., Shiokawa, K., Elsaid, A, AbuElezz, O. A., & Mahrous, A. M. (2023). A machine learning approach for total electron content (tec) prediction over the northern anomaly crest region in egypt [Space and Geophysical Observations and Recent Results related to the African Continent]. *Advances in Space Research*, *72*(3), 790–804. https://doi.org/https://doi.org/10.1016/j.asr.2022.10.052

Chen, Z., Jin, M., Deng, Y., Wang, J.-S., Huang, H., Deng, X., & Huang, C.-M. (2019). Improvement of a deep learning algorithm for total electron content (tec) maps: Image completion. *Journal of Geophysical Research: Space Physics*, *124*. https://doi.org/10.1029/2018JA026167

Sorkhabi, O. M. (2021). Deep learning of total electron content. *SN Applied Sciences*, *3*(7), 685. https://doi.org/10.1007/s42452-021-04674-6

Mengting, Y., Ziming, Z., & Jia, Z. (2020). "an ionospheric tec grid point prediction model" paper data. https://doi.org/https://doi.org/10.12176/01.99.00065

Habarulema, J. B., McKinnell, L.-A., & Cilliers, P. J. (2007). Prediction of global positioning system total electron content using neural networks over south africa. *Journal of Atmospheric and Solar-Terrestrial Physics*, *69*(15), 1842–1850. https://doi.org/https://doi.org/10.1016/j.jastp.2007.09.002

Kumar, S., & Singh, A. (2009). Variation of ionospheric total electron content in indian low latitude region of the equatorial anomaly during may 2007–april 2008. *Advances in Space Research*, *43*(10), 1555–1562.

Patel, N. C., Karia, S. P., & Pathak, K. N. (2017). Gps-tec variation during low to high solar activity period (2010-2014) under the northern crest of indian equatorial ionization anomaly region. *Positioning*, *8*(2), 13–35.

Bagiya, M. S., Joshi, H. P., Iyer, K. N., Aggarwal, M., Ravindran, S., & Pathan, B. M. (2009). Tec variations during low solar activity period (2005ndash;2007) near the equatorial ionospheric anomaly crest region in india. *Annales Geophysicae*, *27*(3), 1047–1057. https://doi.org/10.5194/angeo-27-1047-2009

Olwendo, O., Yamazaki, Y., Cilliers, P., Baki, P, & Doherty, P. (2016). A study on the variability of ionospheric total electron content over the east african low-latitude region and storm time ionospheric variations. *Radio Science*, *51*(9), 1503–1518.

Oryema, B, Jurua, E, D'ujanga, F., & Ssebiyonga, N. (2015). Investigation of tec variations over the magnetic equatorial and equatorial anomaly regions of the african sector. *Advances in Space Research*, *56*(9), 1939–1950.

Li, Q., Hao, Y., Zhang, D., & Xiao, Z. (2018). Nighttime enhancements in the midlatitude ionosphere and their relation to the plasmasphere. *Journal of Geophysical Research: Space Physics*, *123*(9), 7686–7696.