Machine Learning Model Development for Space Weather Forecasting in the Ionosphere

Randa Natras[1], Michael Schmidt[1]

[1]Deutsches Geodätisches Forschungsinstitut der Technischen Universität München (DGFI-TUM), Arcistrasse 21, Munich, 80333, Germany

Abstract

In this paper, the workflow of the machine learning model development for the space weather forecast in the Earth's ionosphere is presented, as an ongoing project. The problem of space weather forecasting using traditional approaches is discussed, as well as the advantages of using machine learning instead. In addition, the methods and approaches for building a machine learning model are presented, together with challenges related to data and algorithms. The machine learning workflow for the problem of space weather forecast is discussed from problem formulation and data acquisition, data preparation and feature engineering, learning algorithms, to model training, evaluation and deployment. This paper provides an overview of a machine learning project for space weather forecasting and discusses challenges and open issues.

Keywords

Machine Learning, Deep Learning, Model Development, Ionosphere, Space Weather Forecast

1. Introduction

Space weather describes conditions caused by the Sun in the near-Earth space, i.e. magnetosphere, ionosphere and thermosphere that can influence the performance and reliability of space-borne and ground-based technological systems. It can produce major disturbances of Earth's magnetosphere known as geomagnetic storms. Numerous effects of strong space weather on satellites, power grids, aviation, communication and navigation systems have already been observed and documented with considerable economic losses [1, 2, 3, 4]. As society increasingly relies on the services that these infrastructures provide, there is an urgent need to develop advanced forecasting capabilities in order to be able to mitigate a catastrophic failure of space- and ground-based technological systems associated with this type of hazard [5]. The impact of space weather on the Earth's ionosphere and GNSS-based (Global Navigation Positioning System) applications can be modelled by quantifying the Vertical Total Electron Content (VTEC) within the ionosphere. The ionosphere represents the ionized region of the upper atmosphere (from about 50 km up to 1,000 km or more from the Earth's surface) that contains free electrons and ions produced by solar radiation [6]. Free electrons in the ionosphere affect the propagation of a microwave signal and induce a delay or advance of the signal. VTEC is proportional to the relative ionospheric delay of GNSS signals, measured in TEC units (TECU), where 1 TECU = 10^{16} electrons/m^2 [7]. During severe space weather conditions, the variability in the ionosphere can increase drastically in time and space. These sudden intense variabilities are often difficult to model with traditional mathematical approaches and to properly minimize in positioning solutions leading to degradation of positioning and navigation performance [8, 9, 10]. Besides the space weather disturbances, the Earth's ionosphere exhibits considerable geographical variations, which can be divided into high-latitude (beyond ±60° geomagnetic latitude), low-latitude (within ±20° of each side of the geomagnetic equator) and mid-latitude zones (between the boundaries of the other two zones) [11]. Other variations in the Earth's ionosphere, important to mention, depend on local time (daily variations), longitude, season and 11-year sunspot cycle. In order to model and predict the space weather, a complex chain of physical processes between the Sun, the interplanetary space, the Earth's magnetic field and the ionosphere have to be taken into account. However, we have limited understanding of these coupled processes and often do not know the physical and/or mathematical relationships to describe them properly. On the other hand, there are numerous data from satellites and observatories that monitor space weather processes between the Sun and the Earth. Artificial Intelligence and Machine Learning (ML) offer a new possibility of learning from data, in contrast to traditional programming, where programs with detailed rules need to be written that explicitly instruct a computer how to execute steps (Figure 1). In the case of space weather forecast, the traditional programming approach would be to analyze space weather properties, as well as typical and space weather-induced variations in the ionosphere to detect patterns, then write a forecasting algorithm consisting of list of rules for each

CDCEO 2021: 1st Workshop on Complex Data Challenges in Earth Observation, November 1, 2021, Virtual Event, QLD, Australia.
randa.natras@tum.de (R. Natras); mg.schmidt@tum.de (M. Schmidt)
0000-0002-9497-7570 (R. Natras)

of the noticed patterns, test the program and iterate these steps until the model performance is satisfactory. However, the problem of space weather is highly complex and our understanding of the underlying processes of space weather is still limited to be able to properly describe the physical and/or mathematical relations using traditional methods. On the other hand, the ML approach offers the possibility of automatically learning rules from the data that map inputs to outputs. The approach of learning directly from the data can lead to discovering the hidden knowledge of relationships within data and to deepening our physical understanding [12]. These can be achieved, for instance, by estimating the importance of the input variables to the output(s) in a ML model, or by finding structures and relationships within the data through unsupervised learning etc. Furthermore, ML can be used to estimate the nonlinear functions that describe the underlying space weather processes [12] based on the data provided. Recently, there has been an increasing interest in the ML applications for forecasting space weather in the ionosphere using Deep Learning (DL) methods such as Feed-forward Neural Network [13], Autoregressive Neural Network [14], Long Short-Term Memory (LSTM) [15]. Some studies used other ML algorithms like Extreme Gradient Boosting (XGBoost) [16] and Gaussian Process Regression [17]. The DL models proved to be more accurate than traditional modelling approaches such as AutoRegressive Integrated Moving Average (ARIMA) and Empirical Orthogonal Functions (EOF) [15, 13].

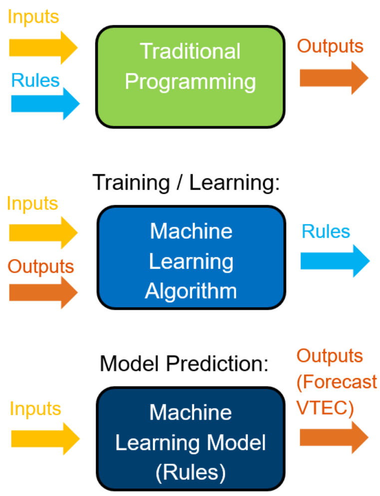

Figure 1: Comparison of the traditional programming approach and ML. Top: Traditional programming approach, where the computer is supplied with input data and an explicitly specified extensive list of rules. Bottom: ML approach, where in the training phase, the computer is supplied with prepared examples of inputs and outputs (training data) and rules are learned from the data. In the next step of the model prediction (test data and model deployment phase), new input data (unseen during the training phase) is fed into the ML model that contains already learned ruled. Results are predicted values, in this case, the forecast of the VTEC in the ionosphere.

2. Machine learning model development

This section presents learning algorithms, data challenges, planned methodologies and approaches to be used in this study. The ML model development workflow for space weather forecast can be summarized into four main phases (Figure 2):

1. Problem formulation and data acquisition
2. Data exploration and feature engineering
3. Model training and cross-validation
4. Model final evaluation (test phase) and deployment of the model.

2.1. Problem formulation and data acquisition

Tim Mitchell [18] proposed a definition of a well-posed ML problem: "A computer program is said to learn from experience E with respect to task T and a performance measure P, if its performance on T, as measured by P improves with experience E". In the context of this study, the ML problem is defined with the task of VTEC forecast, where the experiences are provided in a form of training data and a performance measure is chosen as root mean square error (RMSE) and correlation coefficient. The task of predicting space weather manifestation in the ionosphere is defined via forecasting VTEC in the ionosphere. The problem is formulated in such a way that it can be solved with supervised learning. Supervised learning can be seen as a function approximation or predictive learning problem [19]. Using a training sample of the input (predictors, features or the independent variables) and output (response or the dependent variable) vector, the goal is to obtain an approximation of the function that optimally describes the relationship between input and output. This is achieved by minimizing a certain loss function over the joint distribution of all values. For this study, data of solar activity, solar wind, geomagnetic field are collected from NASA/GSFC through OMNIWeb (https://omniweb.gsfc.nasa.gov/form/dx1.html) [20]. The VTEC values are extracted for high-, mid- and low-latitude points from the GIM (Global Ionosphere Map) provided by the CODE

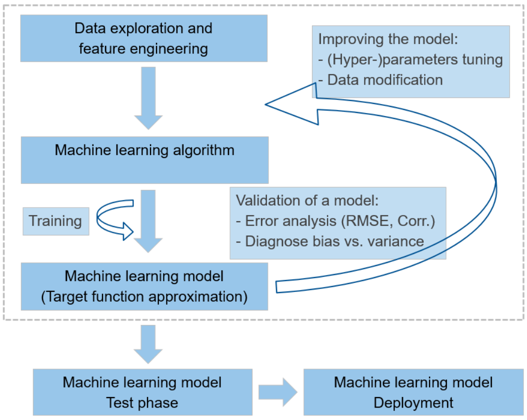

Figure 2: Machine learning model development workflow for the space weather forecast as an iterative process. It starts with the initial data selection, feature extraction and the selection of learning algorithm for model training. Based on the error diagnostic, (hyper-)parameters of the learning algorithm are modified. Since the ML model "learns" from the data, they are modified iteratively until the most suitable features are found that can provide optimal cross-validation results. When the process of the feature selection, algorithm selection and model training phase is complete, the model can be evaluated on previously unseen test data.

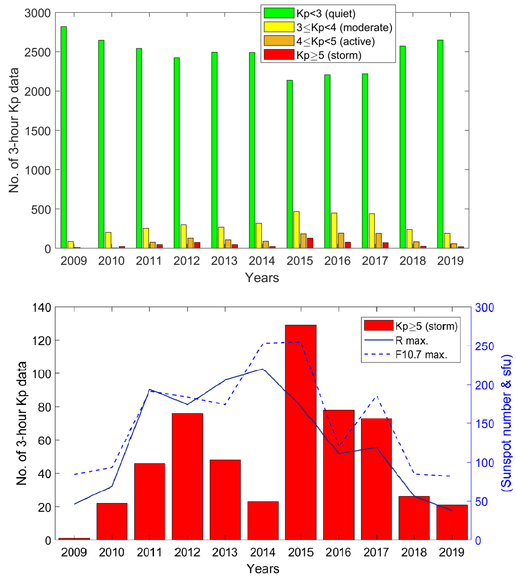

Figure 3: Overview of the geomagnetic activity for the solar cycle 24 (2009 – 2019), based on the geomagnetic index Kp. Top: number of all 3-hour Kp data, bottom: number of 3-hour Kp 5 data that indicate a geomagnetic storm and the maximum values of the sunspot number (R) and the solar radio flux F10.7 as indicators of solar activity.

(https://cddis.nasa.gov/archive/gnss/products/ionex) [21].

2.2. Data exploration and feature engineering

The goal of this step is to identify the relevant predictors and prepare a dataset (preprocessed and cleaned up) that will be useful for the learning task. It is crucial that the training set represents the ultimate task of the model, contains multiple cases and accurately represents the operation data that may be encountered in practice. Feature engineering [22] refers to the process of transforming raw data into suitable features that can better represent the underlying problem, here space weather forecast. It includes various steps such as feature selection, feature extraction, feature scaling, feature transformation etc. However, it can be a challengeable task to select representative dataset and features that can describe all the cases that can occur in practice. This is an iterative process, where an initial dataset is used in the first attempt and according to the performance of the model, the data are iteratively improved (Figure 2). Events of the highest interest are geomagnetic storms. The occurrence of space weather events was analyzed during solar cycle 24 (from 2009 to 2019), using the geomagnetic activity index Kp (Figure 3).

As can be seen, the state in the geomagnetic field is quiet to moderate the most of the time. Most of the storm events occurred in the years after the maximum of the solar cycle (in April 2014). Thus, the initial periods for training and cross-validation are selected to be 2015 and 2016, while the test year is 2017. The selection of useful features for the model is currently in progress. Explanatory data analysis is applied to understand the data by inspecting its distribution, statistical properties, relationships, correlations etc. Figure 4 shows the distribution of VTEC from the training data for three ionospheric points corresponding to the high-, mid- and low-latitude ionosphere. The distribution peak is around 10 TECU for all three studied regions, but the distribution extends further into the higher values than into the lower values.

Based on the result of the analysis, an important aspect to consider is how to deal with an imbalanced dataset, where cases of space weather appear rarely compared to the quiet period. However, these cases are important as they can produce irregular variations in the ionosphere. An imbalance case can present a difficulty for a learning algorithm, which can lead to a biased model towards the majority of cases [23]. Possible solutions of dealing with imbalance cases may include analyses of the individual

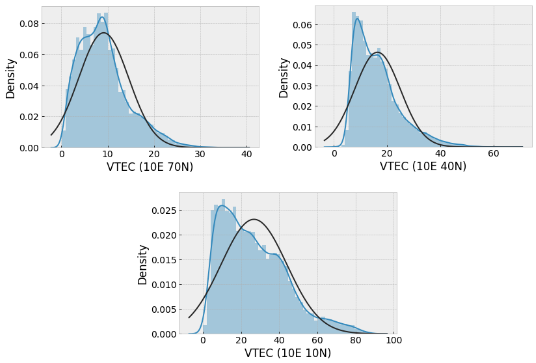

Figure 4: Overview of the distribution of the VTEC training data (2015 – 2016). Gaussian kernel density plot, black curve corresponds to the normal distribution. Top left: VTEC 10°E 70°N, top right: VTEC 10°E 40°N, bottom: VTEC 10°E 10°N.

properties of rare examples in order to distinguish between minority samples and noisy samples, selection of an appropriate learning algorithms and optimal features to enhance learning of rare VTEC signatures, training on the entire and under-sampled datasets, as well as development of cost-sensitive solutions that are able to adapt the penalty with respect to the degree of importance assigned to the minority case.

Another issue with the data is the different temporal and spatial resolution. The temporal resolution of the data covers daily samples (for the solar activity indices R and F10.7), 3-hour data for the geomagnetic index Kp, and hourly to 1-minute samples for other data describing space weather and climate. Data describing solar and geomagnetic activity are given as a function of time, while ionosphere VTEC data are temporally and spatially dependent. VTEC data from the GIM CODE are provided with a temporal resolution of 2 hours until the year 2015 and 1 hour onwards, while spatial sampling is 2.5° x 5° in latitude and longitude. It is important to take into account that GNSS stations used to estimate the GIM are unevenly distributed globally with the current lack of GNSS ground receivers, particularly over the oceans and in the southern hemisphere, among other regions, where most part of the provided VTEC information is based on interpolation, which may result in lower accuracy in these regions.

2.3. Model training

The next step in the ML model development for VTEC forecast is to decide which algorithms, hyperparameters (parameters of learning algorithm) and model architecture should be used. This is done by training an initial model and performing model diagnostic through error analysis. This approach provides insights into the models' performance and gives guidance on how to improve the model. Time series cross-validation [17] is used to evaluate model performance preserving the temporal structure of time series, and to diagnose the bias/variance problem. Overfitting (high variance) lead to very low training errors, but high validation and test errors, while underfitting (high bias) leads to high errors in all the datasets. The aim of this step is to identify the right complexity for a model in order to avoid both underfitting and overfitting. The complexity of the model is changed by altering various hyperparameters of the learning algorithm, increasing/decreasing regularization, getting more training data, adding or removing features, etc. [18]. Which step should be taken depends on what we want to fix: bias (to increase complexity) or variance (to decrease complexity) of the model. One way to diagnose this problem is to plot the learning curves for training and cross-validation datasets. Another important issues to address are interpretability and explanability. ML models based on ensemble learning such as Random forest [24] and Boosting [25] provide a possibility to inspect which predictors have been used most often by a learning algorithm. This information can be useful in interpreting the model and understanding the problem of space weather forecasting. In addition, ensemble learning is recognized as method that can provide a significant improvement in robustness to skewed distribution and good predictive power [23]. On the other hand, Artificial Neural Networks (ANN), a core of DL and state-of-the art techniques for many applications nowadays, are often difficult to explain. They require many parameters, which consequently needs careful design in order to not overfit the data [26]. Modelling of spatial-temporal dependencies is another important task in space weather forecasting. DL provides opportunity to automatically extract features in the spatial domain (e.g. Convolutional Neural Networks) and in the temporal domain (e.g. Recurrent Neural Networks), therefore some researchers propose their combination to learn spatial-temporal features ([27, 28, 29]. Other approaches such as decomposing time series into components that capture trend and seasonality [30] and detrending the time series [31] may be a suitable adaptation for other ML / DL algorithms. In Figure 5, time information consisting of the hour of day and day of year are used as inputs to be able to model daily and seasonal VTEC variations. The next important issue to focus on is estimation of the uncertainty associated with predictions. Uncertainty can be quantified, for instance, by learning the probability distribution over weights in the ANN [32].

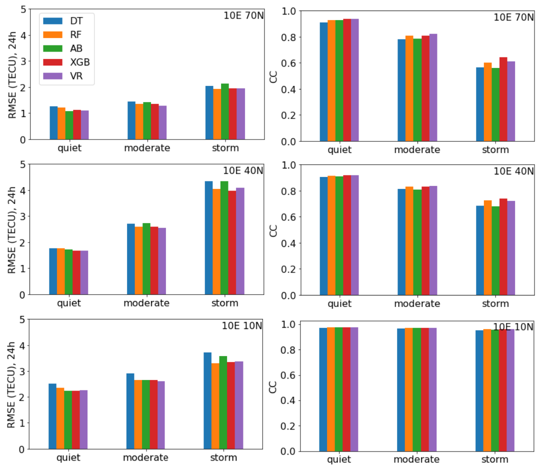

Figure 5: RMSE (left) and correlation coefficients (right) for 1-day forecast for quiet (Kp<3), moderate (3Kp<5) and storm (Kp5) periods during test year 2017 for Decision Tree (DT), Random Forest (RF), AdaBoost (AB), Optimized Gradient Boosting (XGBoost) and Voting Regressor (combines RF, AB and XGBoost). Top: VTEC 10°E 70°N, middle: VTEC 10°E 40°N, bottom: VTEC 10°E 10°N.

2.4. Model final evaluation and model deployment

The generalization error is estimated using a test dataset to show how accurately the model can predict outcome values for previously unseen data. Figure 5 shows the experimental results for the 24-hour forecast with the decision tree and ensemble learning for the quite, moderate and storm periods of the year 2017, based on data of solar activity, solar wind, geomagnetic field, time information (hour of day and day of year) and ionosphere VTEC [33]. The RMSE for the storm period is up to two times higher than for the quiet period, while the correlation coefficient decreases as the Kp index increases. Ensemble learning methods achieve better accuracy than a single decision tree, where combining multiple ensembles gives the most optimal results.

Further exploration of the data and algorithms will be carried out to improve learning during the storm and to address the questions raised in Section 2.2 and Section 2.3. The ultimate goal is to build a model that generalizes the problem well and fits the data reasonably well. In the model deployment phase (application of the model in practice) important components will be monitoring and maintaining the model by tracking various metrics.

3. Conclusion

Many complex physical problems, such as space weather forecasting, have challenges to be predicted accurately and over a longer term. One of the great challenges for space weather are the variable dynamical processes between the Sun and the Earth for which there is still not enough understanding of all underlying processes and relationships. Accurate forecasting and early-warning systems are urgently needed for today's society, which relies on space- and ground-based technological infrastructures, which are particularly vulnerable to extreme space weather events and not adequately protected. The ML approach has the possibility to find nonlinear functions to approximate space weather processes and forecast their manifestation in the Earth's magnetic field and in the ionosphere. This paper introduces the workflow of ML model development for the forecasting of space weather in the ionosphere. The study is still in progress. The models are data-driven, gaining knowledge from data, which describe solar activity, solar wind speed, geomagnetic field and the ionosphere. To get the most out of ML, it is required to find and prepare relevant data and useful features that can enhance learning, especially during space weather events. When facing an imbalance dataset, an intelligent system must be developed that is able to overcome such bias. In addition, the uncertainty of the weather forecast should be provided. It is also important to understand and be able to explain what the model has learned and whether it is biased in any way. Inspecting what the ML model has learned can also help us to discover hidden patterns and to gain a better understanding of the problem.

Acknowledgments

This work is funded by Research Grants - Doctoral Programmes in Germany from German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD).

References

[1] E. Krausmann, E. Andersson, M. Gibbs, W. Murtagh, Space weather and critical infrastructures: Findings and outlook, EUR 28237 EN (2016).

[2] L. J. Lanzerotti, Space weather: Historical and contemporary perspectives, Space Science Reviews 212 (2017) 1253–1270. doi:10.1007/s11214-017-0408-y.

[3] D. N. Baker, E. Daly, I. Daglis, J. G. Kappenman, M. Panasyuk, Effects of space weather on technology infrastructure, Space Weather 2 (2004). doi:10.1029/2003SW000044.

[4] J. P. Eastwood, E. Biffis, M. A. Hapgood, L. Green, M. M. Bisi, R. D. Bentley, R. Wicks, L.-A. McKinnell, M. Gibbs, C. Burnett, The economic impact of space weather: Where do we stand?, Risk Analysis 37 (2017) 206–218. doi:10.1111/risa.12765.
[5] D. J. Knipp, J. L. Gannon, The 2019 national space weather strategy and action plan and beyond, Space Weather 17 (2019) 794–795. doi:10.1029/2019SW002254.
[6] A. D. Richmond, Ionosphere, Springer Netherlands, Dordrecht, 2007, pp. 452–454. doi:10.1007/978-1-4020-4423-6_159.
[7] B. Hofmann-Wellenhof, H. Lichtenegger, J. Collins, Global Positioning System: Theory and Practice, Springer, Berlin, 2001. doi:10.1007/978-3-7091-6199-9.
[8] M. Poniatowski, G. Nykiel, Degradation of kinematic ppp of gnss stations in central europe caused by medium-scale traveling ionospheric disturbances during the st. patrick's day 2015 geomagnetic storm, Remote Sensing 12 (2020). URL: https://www.mdpi.com/2072-4292/12/21/3582. doi:10.3390/rs12213582.
[9] I. Zakharenkova, I. Cherniak, Effects of storm-induced equatorial plasma bubbles on gps-based kinematic positioning at equatorial and middle latitudes during the september 7–8, 2017, geomagnetic storm, GPS Solutions 25 (2021). doi:10.1007/s10291-021-01166-3.
[10] X. Luo, S. Gu, Y. Lou, C. Xiong, B. Chen, X. Jin, Assessing the performance of gps precise point positioning under different geomagnetic storm conditions during solar cycle 24, Sensors 18 (2018). URL: https://www.mdpi.com/1424-8220/18/6/1784. doi:10.3390/s18061784.
[11] L. R. Cander, Ionospheric variability, in: Ionospheric space weather, Springer, 2019, pp. 59–93.
[12] E. Camporeale, S. Wing, J. Johnson, Machine learning techniques for space weather, Elsevier, Amsterdam, Netherlands, 2018.
[13] J. C. Uwamahoro, J. B. Habarulema, Modelling total electron content during geomagnetic storm conditions using empirical orthogonal functions and neural networks, Journal of Geophysical Research: Space Physics 120 (2015) 11,000–11,012. doi:10.1002/2015JA021961.
[14] C. Cesaroni, L. Spogli, A. Aragon-Angel, M. Fiocca, V. Dear, G. De Franceschi, V. Romano, Neural network based model for global total electron content forecasting, J. Space Weather Space Clim. 10 (2020) 11. doi:10.1051/swsc/2020013.
[15] R. Tang, F. Zeng, Z. Chen, J.-S. Wang, C.-M. Huang, Z. Wu, The comparison of predicting storm-time ionospheric tec by three methods: Arima, lstm, and seq2seq, Atmosphere 11 (2020). doi:10.3390/atmos11040316.
[16] A. V. Zhukov, Y. V. Yasyukevich, A. E. Bykov, Correction to: Gimli: Global ionospheric total electron content model based on machine learning, GPS Solutions 25 (2021) 21. doi:10.1007/s10291-020-01063-1.
[17] L. Mallika I, D. V. Ratnam, S. Raman, G. Sivavaraprasad, Machine learning algorithm to forecast ionospheric time delays using global navigation satellite system observations, Acta Astronautica 173 (2020) 221–231. doi:10.1016/j.actaastro.2020.04.048.
[18] T. Mitchell, Machine learning (1997).
[19] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference and prediction, 2 ed., Springer, New York, NY, 2009. doi:10.1007/978-0-387-84858-7_3.
[20] J. H. King, N. E. Papitashvili, Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data, Journal of Geophysical Research: Space Physics 110 (2005). doi:10.1029/2004JA010649.
[21] R. Dach, S. Schaer, D. Arnold, E. Orliac, L. Prange, A. Susnik, A. Villiger, A. Jaeggi, Code final product series for the igs (2016).
[22] A. Zheng, A. Casari, Feature engineering for machine learning: principles and techniques for data scientists, O'Reilly Media, Inc., 2018.
[23] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (2016) 221–232. doi:10.1007/s13748-016-0094-0.
[24] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.
[25] J. H. Friedman, Greedy function approximation: A gradient boosting machine., The Annals of Statistics 29 (2001) 1189 – 1232. doi:10.1214/aos/1013203451.
[26] C. C. Aggarwal, Teaching deep learners to generalize, in: Neural Networks and Deep Learning, Springer, 2018, pp. 169–216. doi:10.1007/978-3-319-94463-0_4.
[27] M. Kaselimi, A. Voulodimos, N. Doulamis, A. Doulamis, D. Delikaraoglou, Deep recurrent neural networks for ionospheric variations estimation using gnss measurements, IEEE Transactions on Geoscience and Remote Sensing (2021) 1–15. doi:10.1109/TGRS.2021.3090856.
[28] H. Wilms, M. Cupelli, A. Monti, T. Gross, Exploiting spatio-temporal dependencies for rnn-based wind power forecasts, in: 2019 IEEE PES GTD Grand International Conference and Exposition Asia (GTD Asia), 2019, pp. 921–926. doi:10.1109/GTDAsia.2019.8715887.
[29] L. Mou, L. Bruzzone, X. X. Zhu, Learning spectral-

spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery, IEEE Transactions on Geoscience and Remote Sensing 57 (2019) 924–935. doi:10.1109/TGRS.2018.2863224.

[30] R. Asadi, A. C. Regan, A spatio-temporal decomposition based deep neural network for time series forecasting, Applied Soft Computing 87 (2020) 105963. doi:10.1016/j.asoc.2019.105963.

[31] F. Montesino Pouzols, A. Lendasse, Effect of different detrending approaches on computational intelligence models of time series, in: The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–8. doi:10.1109/IJCNN.2010.5596314.

[32] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297. doi:10.1016/j.inffus.2021.05.008.

[33] R. Natras, M. Schmidt, Ionospheric vtec forecasting using machine learning, in: EGU General Assembly Conference Abstracts, 2021, pp. EGU21–8907. doi:10.5194/egusphere-egu21-8907.