Published in *Proceedings of the IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe) 2022*, Novi Sad, Serbia, 10-12 october 2022, which should be cited to refer to this work.
DOI: 10.1109/ISGT-Europe54678.2022.9960543

A Data-Driven Algorithm for Short-Term Prediction of Over-Voltage and Under-Voltage Events in Distribution Grids

Shabnam Ataee
School of Engineering and Management Vaud (HEIG-VD)
University of Applied Sciences of Western Switzerland (HES-SO)
Yverdon-les-Bains, Switzerland
shabnam.ataee@heig-vd.ch

Mohammad Rayati
School of Engineering and Management Vaud (HEIG-VD)
University of Applied Sciences of Western Switzerland (HES-SO)
Yverdon-les-Bains, Switzerland
mohammad.rayati@heig-vd.ch

Carlos Andrés Pena
School of Engineering and Management Vaud (HEIG-VD)
University of Applied Sciences of Western Switzerland (HES-SO)
Yverdon-les-Bains, Switzerland
carlos.pena@heig-vd.ch

Omid Alizadeh-Mousavi
Depsys SA
Puidoux, Switzerland
omid.mousavi@depsys.com

Mokhtar Bozorg
School of Engineering and Management Vaud (HEIG-VD)
University of Applied Sciences of Western Switzerland (HES-SO)
Yverdon-les-Bains, Switzerland
mokhtar.bozorg@heig-vd.ch

***Abstract*— This paper proposes algorithms for short-term over- and under-voltage prediction in distribution grids. The proposed algorithms are developed using time-series of voltage and current measurements, which does not require the knowledge of distribution grid model (topology and parameters of the components). Various algorithms based on random forest classifier (RFC) and random forest regressor (RFR) methods, two prominent machine learning methods, are developed regarding different feature selection possibilities. The developed algorithms are tested and validated on two real datasets (grid measurement data from GridEye devices in two low voltage grids in Switzerland). An algorithm based on RFR method, with recent information including the measurement data of the last week at the same time of prediction, outperforms other algorithms. The proposed algorithm can predict over- and under-voltage events with 85% accuracy four hours ahead of the real time.**

Keywords*— *Distribution grid, machine learning, random forest classifier (RFC), random forest regressor (RFR), voltage events prediction.

I. Introduction

A. Context and Motivation

To achieve the goals of the energy transition toward a low-carbon future, Distribution System Operators (DSOs) must take an active role in the future of electrical grids to accommodate distributed resources. In particular, the penetration of photovoltaic (PV) systems and charging stations for electric vehicles (EVs) is expected to grow at a rapid pace in the next few years. In this context, distribution grids' observability and controllability have become increasingly valuable to ensure a secure and efficient operation of the grid in the presence of such resources.

Many new technologies have recently been developed to improve the observability of electrical distribution grids in a cost-effective manner. Refs. [1–3] have investigated the advantages and applications of micro-phasor measurement units (u-PMUs) in distribution grids. Nevertheless, medium and low voltage distribution grids are composed by a larger number of nodes and lines compared to the transmission networks. Hence, for economic and technical reasons, the DSOs are looking for affordable and scalable solutions for the network supervision with limited number of measurement devices [4]. As an example, GridEye devices and monitoring system [5] have been developed for the same purpose to: (i) measure voltage, current, active power, and reactive power, and (ii) collect large datasets in distribution grids and make it possible to execute data-driven algorithms.

The controllability of distribution grids, on the other hand, does not become completely automatic yet as it is less cost-justifiable. Most of the time, DSOs control the state of switches, tap positions of transformers, and, if possible, shed the loads in an offline manner [6]. Face the fact that as the penetration of PV systems in distribution grids increases, over- and under-voltage events become more frequent. If the DSOs could predict such events, say, a few hours in advance, they would take the necessary corrective actions to prevent the grid operational limit violations.

By taking advantage of increased observability and collecting data at various measuring points, we could now consider implementing a data-driven algorithm based on machine learning or deep learning for the detection of over- and under-voltage events in distribution grids. We have historical data on voltage, current, active power, and reactive power at various points on a low voltage distribution grid and are attempting to build a short-term predictive model for over- and under-voltage events. The main challenge here is identifying relevant features from the collected data to use in the model. In order to find a concise set of features, we must conduct exploratory data analysis to determine the correlations between variables. Prior to that, we must run a data cleaning pipeline to deal with missing data.

In this paper, we address the aforementioned problem by proposing a pipeline that includes data preparation, missing value and outlier removal, exploratory data analysis, machine learning model implementation, and post-modeling analysis. Finally, short-term prediction models based on machine learning, i.e., random forest classifier (RFC) and random forest regressor (RFR), and deep learning are compared.

B. Literature Review

We survey previous studies on short-term over- and under-voltage event prediction, from two perspectives: (i) the power system point of view, and (ii) the machine learning point of view.

From the standpoint of power systems, many studies, for example, [7–8], have investigated automatic voltage control in distribution grids to avoid over- and under-voltage events. The majority of algorithms that have been developed are based on decentralized control. The cost justification for such systems in distribution grids, however, has been overlooked. The short-term over- and under-voltage event prediction is required for DSOs to implement an offline control mechanism by taking preventive actions. In [9], an algorithm for calculating voltage sensitivity factors to active or reactive power production and consumption of distribution grid nodes was proposed. An indirect approach was used to first predict related variables such as active and reactive power of nodes (production and consumption). The voltages were then calculated as a function of the active and reactive power of the nodes based on the sensitivity coefficients. In [10], nodal voltages are controlled in distribution grids with high PV system penetration using historical active and reactive power measured at different nodes and sensitivity factors.

From the standpoint of machine learning, short-term over- and under-voltage event prediction is a time-series classification problem. To solve the time-series prediction problems such as weather forecasting, future traffic prediction, stock market price trend prediction, and so on, various machine learning and artificial intelligence algorithms are typically used. For short-term and long-term prediction in various time-series problems, methods based on K-nearest neighbors [11], support vector machine (SVM) [12], decision tree [11], random forest [11], and deep learning models such as convolutional neural networks (CNN) [13] and long short-term memory (LSTM) [14] have been proposed in the literature. Another aspect of the problem we face is the importance of feature selection. Feature selection is an effective step before applying machine learning models, which reduce the computation time, improve learning accuracy, and facilitate a better understanding of the learning model or data by removing irrelevant and redundant features [15].

C. Contribution and Novelty

In this paper, over- and under-voltage events are directly predicted using the historical measurement data. After collecting measurement data in distribution grids using cost-effective technologies, we developed a data-driven algorithm based on RFC and RFR for direct shot-term prediction of over- and under-voltage events. It should be noted that short-term prediction of active and reactive power at each node of the distribution grid with high accuracy is difficult due to the variability and heterogeneous nature of nodal load profiles. Overall, the direct short-term voltage prediction compared to the indirect one outperforms because the voltage variations are less than the active, reactive power, and current variations in distribution grids.

The followings are the study's main contributions:

- A pipeline is formed for developing a machine learning model that predicts over- and under-voltage events four hours in advance.
- A feature selection strategy is proposed for training the proposed machine learning algorithm.

D. Paper Organization

In Section II of this paper, we briefly describe the available datasets. Section III describes the proposed algorithms, which are based on the RFC and RFR methods. The numerical results and related discussions are presented in Section IV. Finally, Section V concludes the paper.

II. Data

Two datasets are used in this paper, collecting by a DSO in Switzerland with GridEye devices. The datasets are for two different distribution grids located in the western Switzerland. Due to confidentiality concerns, the entire datasets cannot be shared. Here, we present the test on one of the dataset.

The considered dataset is composed of 64 nodes, where the data of 49 different nodes are analyzed and cleaned they do not have any non-systematic noises. For each node n, the values of voltage (V_n), active power (P_n), reactive power (Q_n), and the values of current (I_l), active power (P_l), and reactive power (Q_l) of its upward line l for three different phases are available every 10 minutes for a maximum of one year.

Non-systematic components such as noise are not predictable in this paper because we are attempting to solve a short-term time-series prediction problem. Following that, the data from a node (i.e., node 36) with the least noisy voltage time-series is chosen for further investigation. The proposed algorithms are then tested on other nodes.

III. Proposed Algorithm

In the following, we present our algorithm for predicting over- and under-voltage events four hours ahead based on machine learning models tested on node 36 of the considered distribution grid. To solve this classification problem, we followed a pipeline from (A) data preparation; (B) missing values removal; (C) outliers removal; (D) exploratory data analysis; (E) machine learning model implementation; and finally (F) post-modeling analysis. In the following, we will explain steps (A)-(E) in more detail. The post-modeling analysis, i.e., step (F) is explained in the next section.

A. Data Preparation

1) Input features:

The input features are several time-series of a specific node and flowing active and reactive power of its upward line every 10 minutes for three different phases during the one-year period from September 3, 2018, to August 31, 2019. The size of the data is around 52 Kbytes. For the sake of simplicity in this paper, we treated phases independently. Therefore, among 18 different time-series available for the node under study, four time-series, i.e., the voltage values in V for phase A of the node under study, the current values in kA, the active power in kW, and the reactive power in kVAR for phase A of its upward line, are selected for further usage.

2) Target value:

To predict over- and under-voltage events, we defined two different voltage states: 0 (normal) and 1 (undesired). If the voltage value V_n is in an undesirable state, $V_n > 230 + 5\% \cdot E_t\{V_n\}$ or $V_n < 230 - 5\% \cdot E_t\{V_n\}$. The voltage state in the next four hours is defined as the output feature (target value). We can define the target value in timestamp t as the voltage state in timestamp $t + 24$ because the voltage values are available every 10 minutes. Exploring the voltage time-series, we discovered that the voltage values are in an undesirable state 20% of the time.

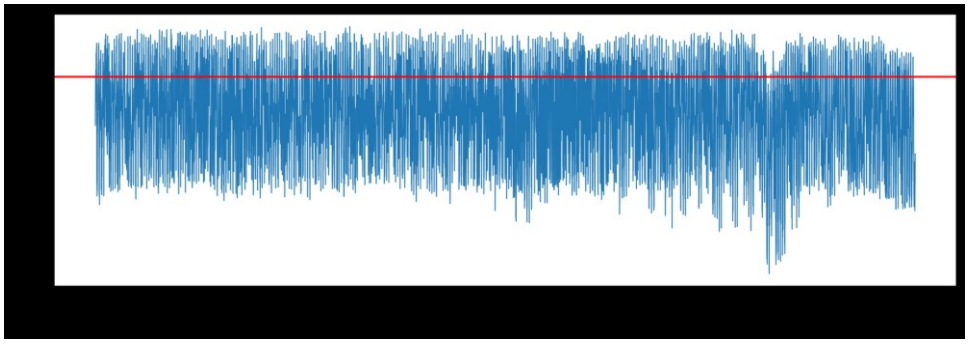

Fig. 1. Voltage values of the first phase in a sepcific node located in the considerd distribution grid.

B. Missing Values Removal

We explored the dataset for missing values and discovered that the measurement device was interrupted once for one hour (on March 31, 2019, at 2:00 AM), once for 30 minutes (on November 30, 2018, at 10:30 AM), and three times for 10 minutes (on April 11 at 9:50 PM, April 12 at 5:50 AM, and July 10 at 2:00 PM). To fill in such gaps, the backward propagation technique is used. The next available value is propagated backward to fill the gap in this technique.

C. Outliers Removal

The voltage values less than $200V$ or more than $300V$ are assumed as *outliers*. We discovered that the voltage time-series contains no outliers.

D. Expolatory Data Analysis

We conducted an exploratory data analysis to become more acquainted with the structure of the data under consideration. The scatter plot with the regression line is shown in Fig. 2 to investigate the relationship between the original features. The plot clearly shows a linear correlation between two time-series of the node under study, namely the current (I) and the active power (P).

Next, we did *hourly*, *weekly*, and *monthly* explorations of the current time-series. In our hourly exploration of the current values (Fig. 3), we discovered that the consumption is higher from 8:00 AM to 8:00 PM, resulting in more current flowing than during the night hours, given that this node is located in a commercial building. As a result, a new feature named "*hour*" is added to the input features to show the dependency of the data on the hour values.

In our weekly exploration of the current time-series, we discovered that the profiles of time-series on working days are not similar to the profiles on weekend days (Fig. 4). On Sundays, the flowing current and active/reactive power are less than on other days. On the other hand, as electricity consumption is higher on working days, the flowing current and active/reactive power are greater. To deal with these differences in the profiles, three Boolean features, named "*isWorkingDay*", "*isSaturday*", and "*isSunday*", are added to the input features for distinguishing the working days from the weekend days.

Finally, monthly exploration of the current values shows the dependency of flowing current and active/reactive power profiles on the months of the year. As shown in Fig. 5, the flowing current is greater in February, whereas the flowing current and active/reactive power are less in September. As a result, a new variable, named as "*month*", is added to the input features to emphasize the dependency of the voltage profile on the month of the year.

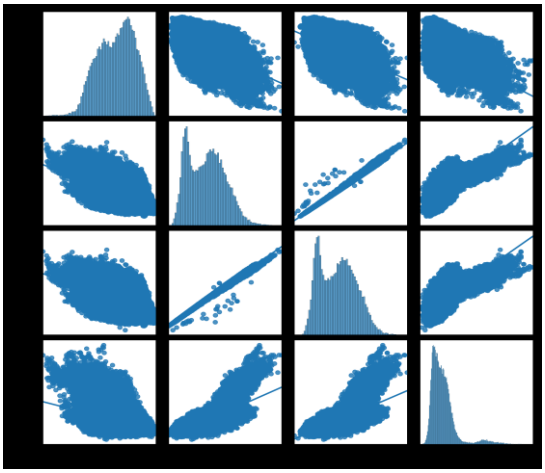

Fig. 2. Correlation between the four original features (V, I, P and Q).

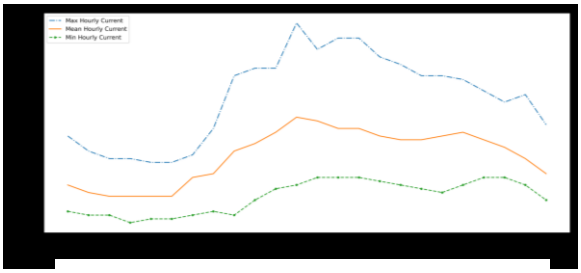

Fig. 3. Hourly expolation of the current time-series.

Further investigation revealed that the time-series profiles of public holidays differ from the time-series profiles of normal working days. To distinguish public holidays from the normal working days, an additional feature, named "*isHoliday*", has been added to the input features, which shows the public holidays in the canton of *Vaud*, Switzerland.

By adding these six features, the number of input features is increased from 4 to 10. In addition, we can add extra temporal features like recent values of voltage, current, and active/reactive power in the last hour, two hours, or last week at the same time. We add them as hyper-parameters to the dataset to analyze and evaluate their impacts on the performance of the proposed machine learning models.

E. Machine learning Model Implementation

Before applying machine learning models, the dataset is divided into a training set (80%), validation set (10%), and test set (10%). The training set is used to fit and train the machine learning models, the validation set is used to tune the algorithm's hyper-parameters, and finally, the test set is used to test and evaluate the trained models. The test set should

Fig. 4. Weekly expolation of the current time-series.

never be used to fit the models and remain unseen before the evaluation.

Two different algorithms are used to solve this classification problem. In the first approach, we looked at the problem as a *time-series classification* problem, while in the second approach, we looked at this problem as a *time-series regression* problem. A *baseline* (based on the first approach) and two machine learning models, i.e., RFC [16] and RFR [17] (based on the first and second approaches, respectively), are fitted on training/validation sets to solve this time-series classification problem. These models are explained in more detail in the following.

1) Baseline (Naïve predictor):

A baseline is a simple algorithm used to create predictions for a dataset. We then compare the performance of any machine learning model with that of the baseline. In the classification problems, the most-frequent category is predicted and used as the baseline. The performance of the most-frequent baseline in this dataset is shown in the following section.

2) Random Forest Classifier (RFC):

A random forest, also known as a random decision forest, is an ensemble algorithm made up of several decision trees that is used to solve classification or regression problems. The random forest is an example of bagging, in which we try to reduce the variance of an estimator, a decision tree in this case, by averaging the predictions from several instances of trained models on different samples of the train set. The random forest is usually used to solve the overfitting problem of the decision tree estimators.

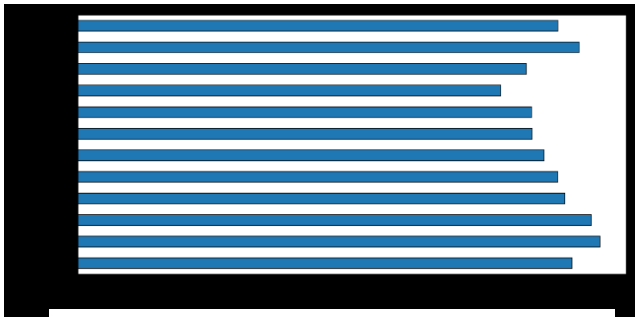

Fig. 5. Monthly expolation of the current time-series.

The grid search is used to the tune hyper-parameters of the random forest classifier, i.e., the number of trees (*n_estimators*) and the maximum depth of each tree (*max_depth*). Five different values for each hyper-parameter are considered, resulting in 25 different combinations of these parameters, in which we evaluate them in a grid search. Then, a random forest classifier is built based on the best value of hyper-parameters to be used for predicting on the test set.

As explained before, we aim to evaluate the impacts of different combinations of extra temporal features on the performance of the adopted machine learning algorithms. The whole process of building and training the model is repeated for each dataset with corresponding input features.

3) Random Forest Regressor (RFR):

The RFR is a random forest estimator used to solve regression problems. This algorithm is composed of two phases: the first is a time-series prediction problem, in which we predict the voltage values in the next four hours using the RFR model. In the second phase, we detect the over- and under-voltage events that will happen in the next four hours.

Like the previous model, we used grid search to tune the hyper-parameters of the estimator, i.e., the number of trees (*n_estimators*) and the maximum depth of each tree (*max_depth*). Three different values for the parameter *n_estimators* and six different values for the parameter max_depth are considered, resulting in 18 different combinations passed to the grid search for evaluation. The best hyper-parameter values are then used to construct a RFR, which is used to predict the test set. The whole process is repeated for different combinations of extra temporal features, which are added to the input features.

IV. Numerical Results and Discussion

As an example, the results of voltage magnitude and over voltage prediction of the proposed RFR algorithm for one node during one week with 10-minutes resolution is depicted in Fig. 6. As can be seen, the green circles represent the moment when the over-voltage was appropriately identified four hours ahead. The red circles, on the other hand, depict the time when the over-voltage event occurred, but the algorithm was unable to detect it four hours ahead.

Two different evaluation metrics, i.e., *accuracy* and *recall,* are used to evaluate the performance of the classification problem in this paper. Accuracy is defined as the fraction of predictions that the model gets correct (either *true positive (tp)* or *true negative (tn))* out of all predictions, i.e.,

$$accuracy = (tp + tn) / (tp + tn + fp + fn),$$

where fp is *false positive* and fn is *false negative*.

Recall is defined as the ratio that the model finds all the positive results, i.e., $recall = tp/(tp + fn)$.

The evaluation results on the test set, using the proposed three algorithms, are given in Table 1. Even though the baseline has a high accuracy (80%), it is not capable of finding any positive samples (*recall* = 0). We observed that all the proposed algorithms outperform the baseline one in terms of higher accuracy and recall.

Among the proposed algorithms, the ones based on RFC achieve higher *accuracy*, whereas the ones based on RFR achieve higher *recall*. The highest accuracy (i.e., 86%) is achieved with the RFC model, where in addition to four original features (V, I, P, and Q) and six overall time information features, four extra recent temporal features, i.e., the values of V, I, P, and Q from yesterday at the same time, are also added to the input features. The recall of this algorithm is 53%.

The highest *recall* (i.e., 84%) is achieved with RFR model, where in addition to 10 original/overall time information features, 24 extra recent temporal features, i.e., the values of V, I, P, and Q for the last hour (6 timestamps) are also added to the input features. The accuracy of this model is 80%.

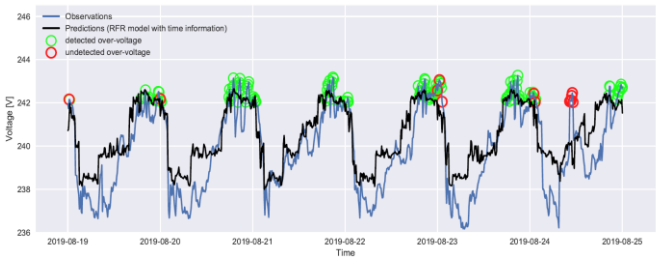

Fig. 6. Predicted and observed over-voltage events.

TABLE I. EVALUATION RESULTS

Alg.	Extra features [a]	Accuracy	Recall
Baseline	---	0.80	0
RFC	---	0.85	0.49
	one hour ago (6 timestamps)	0.84	0.41
	two hours ago (12 timestamps)	0.84	0.41
	yesterday at the same time	0.86	0.53
	last week at the same time	0.86	0.50
RFR	---	0.82	0.77
	one hour ago (6 timestamps)	0.80	0.84
	two hours ago (12 timestamps)	0.82	0.73
	yesterday at the same time	0.83	0.79
	last week at the same time	0.85	0.66

[a] Extra features contain some recent time information.

Among all the proposed models, the RFR model, when we used the data from yesterday at the same time as extra recent temporal features, has the best trade-off of *accuracy* (i.e., 83%) and *recall* (i.e., 79%).

In addition to machine learning models, two algorithms based on deep learning models, i.e., 1-D Convolutional Neural Networks (1D-CNN) [18] and a combination of 1D-CNN and Gated Recurrent Units (GRU), which is a variant of Recurrent Neural Networks (RNN) [19], are implemented to predict over- and under-voltage events in the next four hours in the time-series under study. Because the algorithms based on RFC and RFR have better accuracy and recall than the ones based on deep learning for our dataset, we do not present the details of the deep learning models.

V. CONCLUSIONS

Direct short-term prediction of over- and under-voltage events in distribution grids using in-field measurement devices is possible with data-driven machine learning methods such as random forest classifier (RFC) and random forest regressor (RFR). The best modeling method is RFR, which includes extra overall and recent temporal features. Before determining the over- and under-voltage events, the RFR model predicts the voltage values for the next few hours. We use a pipeline to implement such an algorithm, which includes (A) data preparation; (B) missing values removal; (C) outliers removal; (D) exploratory data analysis; (E) machine learning model implementation; and (F) post-modeling analysis. We predict over- and under-voltage with 83 percent accuracy and 79 percent recall using the proposed RFR model.

ACKNOWLEDGMENT

This work is part of the project “Grid Data Digger (GDD): Grid Data Digger: An automated Distribution Grid Operation Assistance Tool using Data-Driven solutions on a big-data Platform”, project number 34775.1 IP-EE, which was supported by Innosuisse.

REFERENCES

[1] Dusabimana, Emile, and Sung-Guk Yoon. "A survey on the micro-phasor measurement unit in distribution networks." Electronics 9, no. 2 (2020): 305.

[2] Pignati, Marco, Lorenzo Zanni, Paolo Romano, Rachid Cherkaoui, and Mario Paolone. "Fault detection and faulted line identification in active distribution networks using synchrophasors-based real-time state estimation." IEEE Transactions on Power Delivery 32, no. 1 (2016): 381-392.

[3] Shahsavari, Alireza, Mohammad Farajollahi, Emma Stewart, Ciaran Roberts, Fady Megala, Lilliana Alvarez, Ed Cortez, and Hamed Mohsenian-Rad. "Autopsy on active distribution networks: A data-driven fault analysis using micro-PMU data." In 2017 North American Power Symposium (NAPS), pp. 1-7. IEEE, 2017.

[4] Paruta, Paola, Thomas Pidancier, Mokhtar Bozorg, and Mauro Carpita. "Greedy placement of measurement devices on distribution grids based on enhanced distflow state estimation." Sustainable Energy, Grids and Networks 26 (2021): 100433.

[5] Carpita, M., A. Dassatti, M. Bozorg, J. Jaton, S. Reynaud, and O. A. Mousavi. "Low voltage grid monitoring and control enhancement: The grideye solution." In 2019 International Conference on Clean Electrical Power (ICCEP), pp. 94-99. IEEE, 2019.

[6] Majumdar, Ankur, Sotirios Dimitrakopoulos, and Omid Alizadeh-Mousavi. "Grid monitoring for efficient flexibility provision in distribution grids." In CIRED 2020 Berlin Workshop (CIRED 2020), vol. 2020, pp. 703-706. IET, 2020.

[7] Bahramipanah, Maryam, Rachid Cherkaoui, and Mario Paolone. "Decentralized voltage control of clustered active distribution network by means of energy storage systems." Electric Power Systems Research 136 (2016): 370-382.

[8] Gupta, Rahul, Fabrizio Sossan, and Mario Paolone. "Grid-aware distributed model predictive control of heterogeneous resources in a distribution network: Theory and experimental validation." IEEE Transactions on Energy Conversion 36, no. 2 (2020): 1392-1402.

[9] Bozorg, Mokhtar, Omid Alizader-Mousavi, Sebastien Wasterlain, and Mauro Carpita. "Model-less/Measurement-based computation of voltage sensitivities in unbalanced electrical distribution networks: experimental validation." In 2019 21st European Conference on Power Electronics and Applications (EPE'19 ECCE Europe), pp. P-1. IEEE, 2019.

[10] Christakou, Konstantina, Mario Paolone, and Ali Abur. "Voltage control in active distribution networks under uncertainty in the system model: A robust optimization approach." IEEE Transactions on Smart Grid 9, no. 6 (2017): 5631-5642.

[11] I. Kumar, K. Dogra, C. Utreja and P. Yadav, "A Comparative Study of Supervised Machine learning Algorithms for Stock Market Trend Prediction," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1003-1007, doi: 10.1109/ICICCT.2018.8473214.

[12] N. I. Sapankevych and R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey," in *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24-38, May 2009, doi: 10.1109/MCI.2009.932254.

[13] N. Xue, I. Triguero, G. P. Figueredo and D. Landa-Silva, "Evolving Deep CNN-LSTMs for Inventory Time Series Prediction," *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2019, pp. 1517-1524, doi: 10.1109/CEC.2019.8789957.

[14] Zahra Karevan, Johan A.K. Suykens, “Transductive LSTM for time-series prediction: An application to weather forecasting”, Neural Networks, Volume 125, 2020, Pages 1-9.

[15] Cai, Jie, Jiawei Luo, Shulin Wang and Sheng Yang. “Feature selection in machine learning: A new perspective.” *Neurocomputing* 300 (2018): 70-79.

[16] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees.* Routledge, 2017.

[17] Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.

[18] Kiranyaz, Serkan, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. "1D convolutional neural networks and applications: A survey." *Mechanical systems and signal processing* 151 (2021): 107398.

[19] Medsker, Larry, and Lakhmi C. Jain, eds. *Recurrent neural networks: design and applications*. CRC press, 1999.