Forecasting of short-term PV production in energy communities through Machine Learning and Deep Learning algorithms

Nikos Dimitropoulos
Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens
Athens, Greece
ndimitropoulos@epu.ntua.gr

Nikolaos Sofias
HOLISTIC S.A.
Athens, Greece
nsofias@holisticsa.gr

Panagiotis Kapsalis
Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens
Athens, Greece
pkapsalis@epu.ntua.gr

Zoi Mylona
HOLISTIC S.A.
Athens, Greece
zmylona@holisticsa.gr

Vangelis Marinakis
Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens
Athens, Greece
vmarinakis@epu.ntua.gr

Niccolo Primo
Coopérnico
Lisboa, Portugal
nprimo@coopernico.org

Haris Doukas
Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens
Athens, Greece
h_doukas@epu.ntua.g

Abstract: Photovoltaic (PV) modules and solar plants are one of the main drivers towards zero-carbon future. Energy communities that are engaging citizens through collective energy actions can reinforce positive social norms and support the energy transition. Furthermore, by incorporating Artificial Intelligence (AI) techniques, innovative applications can be developed with huge potential, such as supply and demand management, energy efficiency actions, grid operations and maintenance actions. In this context, the scope of this paper is to present an approach for forecasting an energy cooperative's solar plant short term production by using its infrastructure and monitoring system. More specifically, four Machine Learning (ML) and Deep Learning (DL) algorithms are proposed and trained in an operational solar plant producing high accuracy short-term forecasts up to 6 hours. The results can be used for scheduling supply of the energy communities and set the base for more complex applications that require accurate short-term predictions, such as predictive maintenance.

Keywords: Photovoltaic, Energy Prediction, Machine Learning, Deep Learning, Short-term prediction. Artificial Intelligence

I. INTRODUCTION

As consumer-empowerment and community-driven initiatives, energy communities can play a key role for social innovation as they reflect a fundamental shift in citizen's behaviour and their role as a consumer [1]. Engaging citizens through collective energy actions can reinforce positive social norms and support the energy transition. Towards this direction, the Clean Energy Package of the European Commission (EC) recognises and offers an enabling legislative framework for 'Citizen Energy Communities' and 'Renewable Energy Communities' [2].

By 2030, the EU will have to increase renewables to 32% share of the energy supply and in order to reach this binding target, an explicit role for citizens and communities is foreseen [3-4]. This is an important step towards the 'energy democracy', as not only it acknowledged the role of democratically controlled communities in the energy transition but it will also help European citizens to set up their own renewable energy projects and protecting them from the big players of the energy market. Successful renewable energy cooperatives generate positive economic, social and environmental outcomes while accelerating the social and psychological dimensions of the global transition towards clean energy sources [5].

One major source of renewable energy is harvesting solar power through PV solar plants. The technology is becoming more widely used globally and year on year PVs make up a bigger part of the energy mix in the European Union (EU). In 2018, the EU output of PV electricity reached the 127 TWh, amounting to 3.9% of the EU's gross electricity output [6]. The coming decade continued growth is foreseen, mostly driven by increased self-consumption and more rooftop PV installations as a path towards a post-lignite era [7].

In order to maximise efficiency and optimise production for supply and demand [8], applications of Artificial Intelligence (AI) and the use of Machine Learning (ML) and Deep Learning (DL) algorithms are being implemented in several domains. More specifically, these are applications for production [9-10], anomaly detection [11-12] and energy disaggregation [13]. Predicting future energy production values, plays an important role in almost every AI-based application, in both short-term and long-term forecasting horizons, as the accuracy of the prediction is the main factor in several applications such as

978-1-6654-0032-9/21/$31.00 ©2021 IEEE

fault prediction, anomalies detection, load balancing and future performance of energy systems.

Table I presents a moderate review of the state-of-the-art approaches for ML/DL applications related to PV.

TABLE I. ML/DL APPLICATIONS WITH PV MODULES

Source	Application	Features	Algorithms
Mashud Rana et al. [9]	Production forecasting	Power data	NNs, SVR, RF, LSTM and CNN
Javier Huertas Tato et al. [10]	Production forecasting	Weather, power data	Random Forests
Joao Pereira et al. [12]	Maintenance/anomaly detection	Power data	Variational Bi-LSTM
Mahdi Khodayar et al. [13]	Production forecasting, energy disaggregation	Pecan Street dataset & REDD dataset.	LSTM, GRU, CNN
Hyung Keun Ahnet et al. [14]	Production forecasting	Weather, power data	Deep-RNN
Jorge Vicente-Gabriel et al. [15]	PV power prediction & PV maintenance/ anomalies	Weather, power data	LSTM
A.A. du Plessis et al. [16]	Production forecasting Low-level power output dynamics	Weather, power data	FFNN, LSTM GRU
Elizaveta Kharlova et al. [17]	Production forecasting	Weather, power data	Sequence to Sequence deep learning model with Attention
Alessandro Betti et al. [18]	Maintenance/anomaly detection	SCADA data, fault taxonomy	Self-Organising Map (SOM) and KPIs

Utilising relevant infrastructure and sensor-based systems are of major importance in energy efficiency [19], optimising supply and demand and base of energy management [20] of a solar plant system and its capabilities. In this context, the scope of this paper is to present an approach for forecasting an energy cooperative's solar plant short term production by using its infrastructure and monitoring system. More specifically, four ML algorithms are proposed and trained in an operational solar plant producing high accuracy short-term forecasts up to 6 hours.

Apart from the introduction, the paper is structured along three sections. The second section provides an overview of the methodology. The results from the application of the selected ML/DL algorithms are summarised in the third section. Finally, the last section is summarising the key issues that have arisen in this paper.

II. METHODOLOGY

A. Overview

The methodology followed, was to gather relevant data for forecasting energy production from a solar plant and test several ML/DL algorithms to evaluate their accuracy. Time-series forecasting and techniques were used, to test the impact of different time lags and how they contribute in the prediction. These tests were aiming to identify the historical data that an energy community would it need in order to generate an accurate prediction. Finally, the prediction horizon was altered to assess the algorithm's ability to produce accurate forecasts in several time frames.

In the figure below the methodology is described.

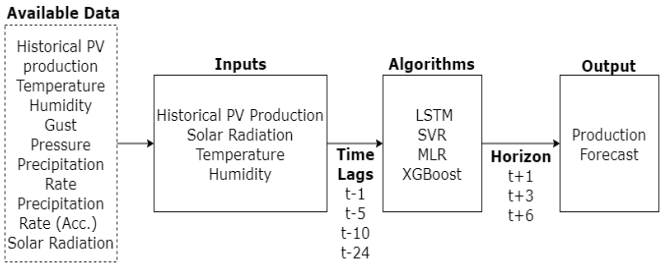

Fig. 1. Methodology schema

The input data would come directly from the solar plant's system, both weather and production data (P_{AC}). However, since the solar plant installation is new (measurements begins in February of 2020), the temperature sensors on site were installed in the early 2021 and to this end, weather data were retrieved from a website that gathers data from a local meteorological station [21] and Copernicus Atmosphere Data Store [22].

The time lags were used with the assumption that the operator would be able to have operational historical data up to 24 hours before. In that way, the architecture of the models offers a timely solution for the required inputs. The prediction horizons are designed for short-term forecasts, with an up to 6 hours horizon.

In this solution, historical data points are solely being used, to eliminate probable errors due to the uncertainty of the weather forecast in future points. The model's output is the predicted hourly P_{AC} production in KWh, in several time horizons.

B. Data Inputs and Pre-processing

The dataset used, consisted of a period spanning over 30 months (August 2018 - January 2021) of production data from a specific solar plant, with a one-hour interval between observations. Data quality of the production data is considered to be very high, as they are deriving from a monitoring platform through the sensor-based system, and the observations of the production are the direct output from 4 DC/AC inverters that are installed.

PV production is related primarily with solar irradiance and the ambient temperature near the site. Since weather data from the on-premises infrastructure is not available, the dataset that the website and the Copernicus database provided, consists of several features. The data quality of the weather information also considered to be high, as main driver of PV production, solar irradiance, is extracted from the Copernicus database [23] [24].

A correlation analysis is used to exclude statistically irrelevant features from the weather variables. As expected, the most statistically relevant parameters are the solar irradiance, temperature, and humidity, which will be used as inputs from

the weather datasets. Table II presents the Pearson and Spearman correlation coefficients of the above variables to the production of the PV are presented, to demonstrate both the linear and non-linear correlations.

TABLE II. CORRELATIONS TABLE

Features	Pearson Coefficient	Spearman Coefficient
Production to Solar Irradiance	0.974	0.9772
Production to Temperature	0.555	0.511
Production to Humidity	-0.6757	-0.6504

C. Algorithms and Models Architecture

Four different algorithms were tested for predicting the future forecasts: Long Short-Term Memory (LSTM), Support Vector Regression (SVR), Multiple Linear Regression (MLR) and XGBoost. In the following table a short description of each algorithm is presented, along with the specific architecture that was used on implementation. The datasets are split in a 80/20 way, meaning that 80% of the data are used to train the models and 20% are used as unseen data to evaluate the algorithms' performance.

TABLE III. ALGORITHMS DESCRIPTION AND ARCHITECTURE

Algorithm	Short Description	Architecture
LSTM	LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. LSTM has feedback connections and can process entire sequences of data. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.	24/48/96/192/284 layers Learning rate: 0.001 Adam optimiser Early stopping patience=5 Epochs: 100 Validation_split: 0.2 Batch_size: 128
SVR	Support-vector machines are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis.	C: 10 Epsilon: 0.01 Timeseries cross validation (4-fold split)
MLR	Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.	Timeseries cross validation (4-fold split)
XGBoost	Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees.	N estimators: 50 Max Depth: 3 Learning Rate: 0.1 Timeseries cross validation (4-fold split)

III. RESULTS AND DISCUSSION

The results are presented below, split by the lagged inputs and the different time horizons, by comparing their R^2 value and their Root Mean Square Error (RMSE), in both train and test sets.

TABLE IV. PREDICTION RESULTS (LAG T-1, HORIZON T+1)

Algorithm	Train set		Test set	
	R^2	*RMSE*	R^2	*RMSE*
LSTM	86.73	1.970	85.96	1.920
SVR	85.89	2.031	84.55	2.014
MLR	86.73	1.970	85.69	1.938
XGBoost	88.56	1.829	**86.66**	**1.871**

TABLE V. PREDICTION RESULTS (LAG T-5, HORIZON T+1)

Algorithm	Train set		Test set	
	R^2	*RMSE*	R^2	*RMSE*
LSTM	96.37	1.030	95.58	1.994
SVR	92.81	1.449	91.41	1.501
MLR	94.68	1.246	93.74	1.282
XGBoost	96.22	1.051	**95.65**	**1.083**

TABLE VI. PREDICTION RESULTS (LAG T-10, HORIZON T+1)

Algorithm	Train set		Test set	
	R^2	*RMSE*	R^2	*RMSE*
LSTM	96.20	1.055	95.15	1.128
SVR	93.43	1.387	91.77	1.471
MLR	94.84	1.229	93.81	1.275
XGBoost	96.39	1.028	**95.56**	**1.081**

The accuracy of these four different algorithms were tested and evaluated in several time lags, 1h, 5h and 10h before the current time for forecasting the next hour's production of the solar plant. The results show that in every algorithm, its accuracy is very high, and their respective RMSEs are very low.

From the four algorithms, the best in each scenario is XGBoost, scoring the best results with an R^2 of 88.56 in the train set and 86.66 in the train set for inputs with one hour lag, and considerably better performance for lagged inputs of 5 and 10 hours, with values 96.22/95.65 of R^2 and 1.051/1.083 of RMSE respectively in the test set.

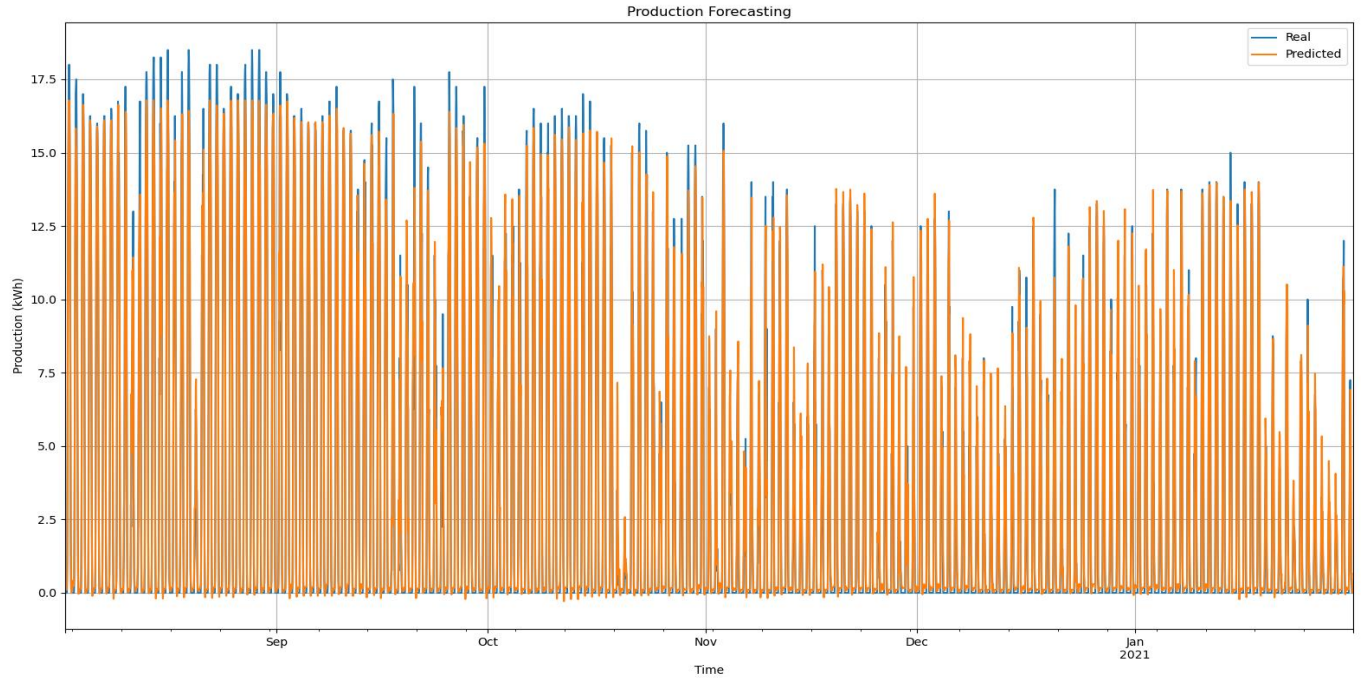

Fig. 2. Predicted and real results plot of XGBoost (lag t-5, horizon 1h)

For the highest performing algorithm, its behaviour on a 24 hours lagged input and a prediction horizon of 3 and 6 hours was tested, in order to assess its performance for forecasts larger than 1 hour. The results are presented below:

TABLE VII. PREDICTION RESULTS, XGBOOST (LAG T-24, HORIZON T+3)

Algorithm	Train set		Test set	
	R^2	*RMSE*	R^2	*RMSE*
XGBoost	92.19	1.509	89.65	1.640

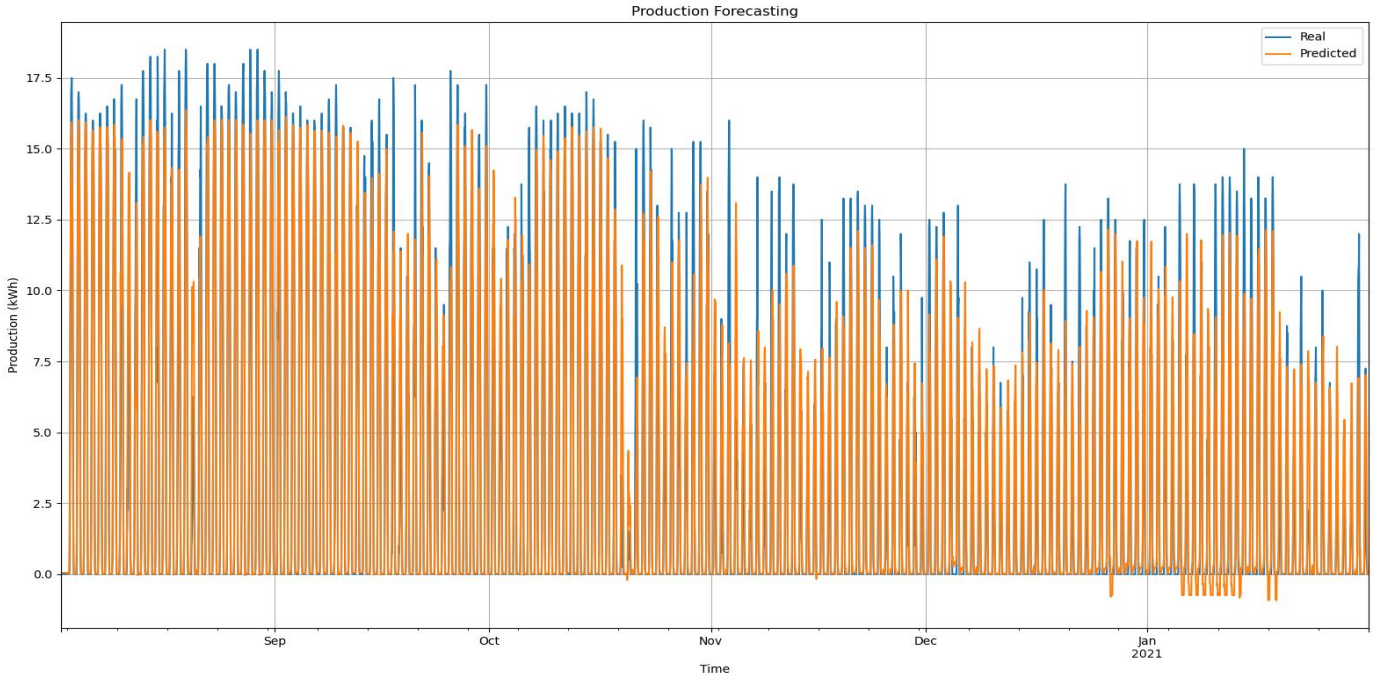

Fig. 3. Predicted and real results plot of XGBoost (lag t-24, horizon 3h)

The accuracy of XGBoost as the highest performing algorithm is still very good, for the 3-hour horizon forecast with a RMSE of 1.509 on the train set and 1.640 on the test set. Although there are several values around zero that the algorithm misses, this could be the fact that there was no optimisation of the algorithm for both the 3- and 6-hours horizon forecasts.

As expected, the 3-hour horizon is better in the respective metrics of the 6-hour one, but still the accuracy of the model that forecast production 6 hours ahead, is still very good, with an RMSE of 1.700 on the train set and 1.834 on the test set.

TABLE VIII. PREDICTION RESULTS, XGBOOST (LAG T-24, HORIZON T+6)

Algorithm	Train set		Test set	
	R^2	*RMSE*	R^2	*RMSE*
XGBoost	90.10	1.700	87.06	1.834

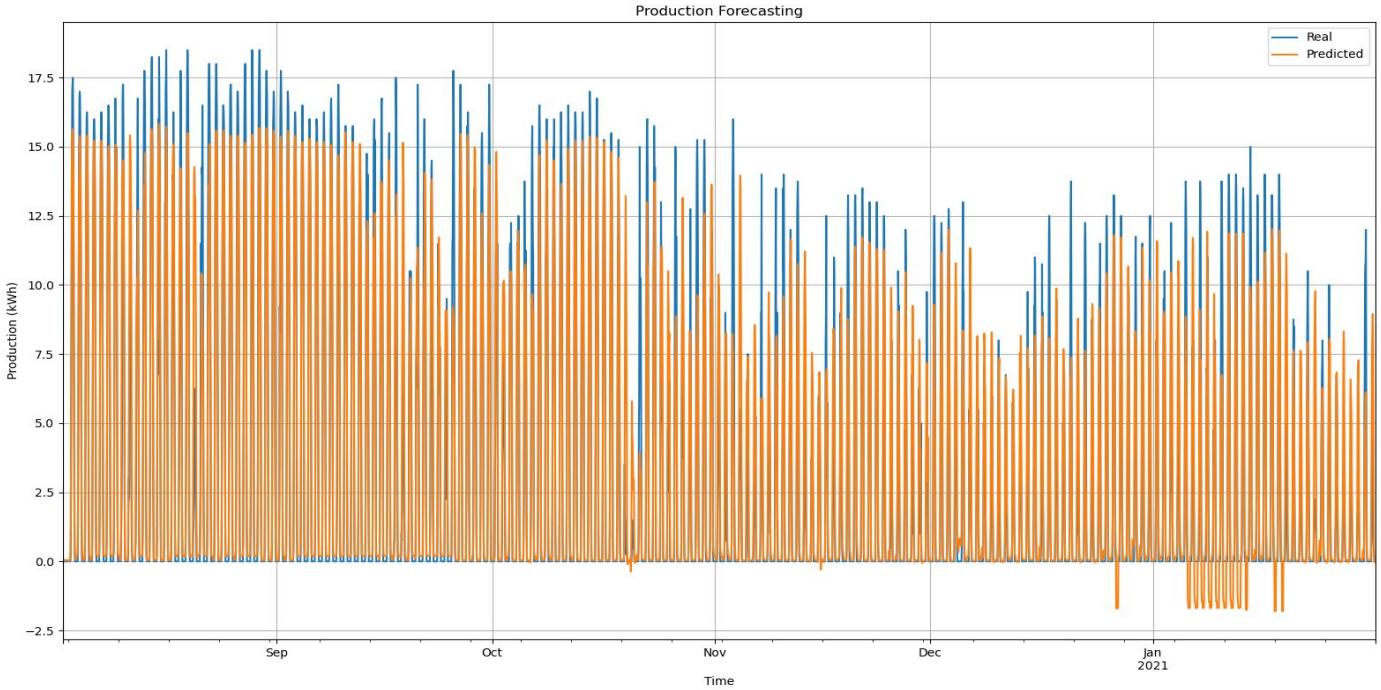

Fig. 4. Predicted and real results plot of XGBoost (lag t-24, horizon 6h)

Finally, the autocorrelation of the production signal itself is calculated, to check how the seasonality of it affects the performance of the algorithms (Fig. 5)

It is observed that there is significant autocorrelation of the production signal, and it is also seasonal. As autocorrelation shows the degree of similarity between a given time series and a lagged version of itself over successive time intervals, this could explain the very similar performances of all algorithms for time lag 5 and 10, as in both of those situations it is observed a high positive autocorrelation.

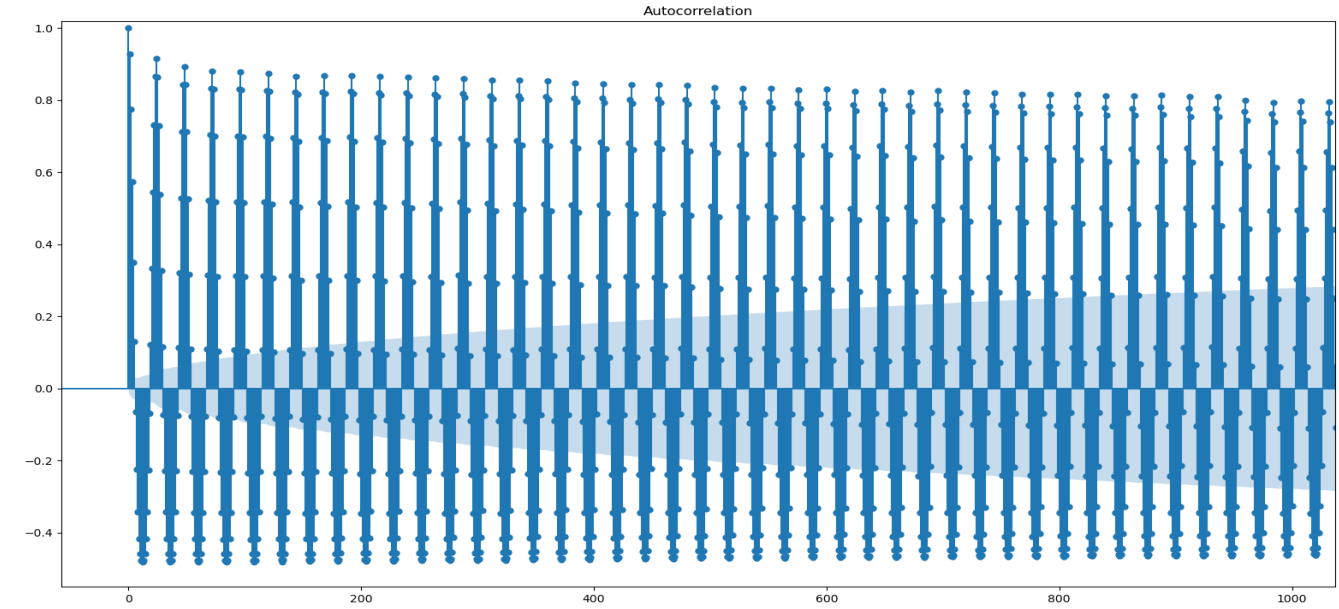

Fig. 5. Autocorrelation plot of production signal

IV. CONCLUSIONS AND NEXT STEPS

In this paper, the energy production of a solar plant's PV modules is forecasted, based only on previous performance and previous weather data. Four different ML/DL algorithms were used for forecasting purposes, LSTM, SVR, MLR and XGBoost and it was found that all the algorithms perform very well, in different combinations of lagged inputs.

The highest performance algorithm was found to be XGBoost, which performed reasonably well in different forecast horizons, thus making it possible to achieve short-term forecasts. The approach followed, requires very few previous observations, in order to make extremely accurate forecast for the next hour, and very good forecasts for the next 3 and 6 hours, demonstrating its efficiency in an energy community environment, where the availability of data is often limited. The results can be used for several action regarding the PVs such as scheduling supply of the energy communities and set the base of more complex applications that require accurate short-term predictions, such as predictive maintenance or energy trading.

The next steps would be to replace the weather variables with the data gathered from the on-site sensors. This will make the predictions even more accurate, since the locality of the weather values plays a key role when predicting the output of PV modules. Furthermore, the autocorrelation of the production signal itself needs to be further investigated as it could possibly drive the forecasts. To this end, energy production forecasting without using weather data is also foreseen so as to investigate the potential and the applicability of different actions regarding the PVs operation.

ACKNOWLEDGMENT

The work presented is based on research conducted within the framework of the project "Modular Big Data Applications for Holistic Energy Services in Buildings (MATRYCS)", of the European Union's Horizon 2020 research and innovation programme under grant agreement no. 1010000158 (https://matrycs.eu/). The content of the paper is the sole responsibility of its authors and does not necessary reflect the views of the EC.

REFERENCES

[1] A. Caramizaru, A. Uihlein, Energy communities: An overview of energy and social innovation, Publications Office of the European Union, Luxembourg, 2020.

[2] European Commission, Directive (EU) 2019/944 on Common Rules for the Internal Market for Electricity and Amending Directive 2012/27/EU & Directive (EU) 2018/2001 on the Promotion of the Use of Energy from Renewable Sources.

[3] S. Soeiro, M. Ferreira Dias, Energy cooperatives in southern European countries: Are they relevant for sustainability targets?, Energy Reports, vol. 6(1), pp. 448-453, 2020.

[4] B. Schmid, T. Meister, B. Klagge, I. Seidl, Energy Cooperatives and Municipalities in Local Energy Governance Arrangements in Switzerland and Germany. The Journal of Environment & Development, 29(1), pp. 123-146, 2020.

[5] M. Tarhan, Renewable Energy Cooperatives: A Review of Demonstrated Impacts and Limitations, Journal of Entrepreneurial and Organizational Diversity, vol. 4(1), pp. 104-120, 2015.

[6] European Commission, Solar power, Available at: https://ec.europa.eu/energy/topics/renewable-energy/solar-power_en (accessed 14/4/2021).

[7] V. Marinakis, A. Flamos, G. Stamtsis, I. Georgizas, Y. Maniatis, H. Doukas, The Efforts towards and Challenges of Greece's Post-Lignite Era: The Case of Megalopolis, Sustainability, vol. 12(24), 10575, 2020.

[8] V. Marinakis, H. Doukas, J. Tsapelas, S. Mouzakitis, Á. Sicilia, L. Madrazo, S. Sgouridis, From big data to smart energy services: An application for intelligent energy management, Future Generation Computer Systems, vol. 110, pp. 572-586, 2020.

[9] M. Rana, A. Rahman, J. Jin, A Data-driven Approach for Forecasting State Level Aggregated Solar Photovoltaic Power Production, 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8.

[10] J. Huertas Tato, M. Centeno Brito, Using Smart Persistence and Random Forests to Predict Photovoltaic Energy Production, Energies, vol. 12, 100.

[11] V. Marinakis, Big Data for Energy Management and Energy-Efficient Buildings, Energies, vol. 13, 1555, 2020.

[12] J. Pereira, M. Silveira. Unsupervised Anomaly Detection in Energy Time Series Data using Variational Recurrent Autoencoders with Attention, 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 1275-1282.

[13] M. Khodayar, M. E. Khodayar, S.M. Jafar Jalali., Deep learning for pattern recognition of photovoltaic energy generation, The Electricity Journal, vol. 34(1), 106882, 2021.

[14] H.K., Ahn, N. Park, Deep RNN-Based Photovoltaic Power Short-Term Forecast Using Power IoT Sensors, Energies, vol. 14, 436, 2021.

[15] J. Vicente-Gabriel, A.-B. Gil-González, A. Luis-Reboredo, P. Chamoso, J.M. Corchado, LSTM Networks for Overcoming the Challenges Associated with Photovoltaic Module Maintenance in Smart Cities, Electronics, vol. 10(1), 78, 2021.

[16] A.A. du Plessis, J.M. Strauss, A.J. Rix. Short-term solar power forecasting: Investigating the ability of deep learning models to capture low-level utility-scale Photovoltaic system behaviour, Applied Energy, vol. 285, 116395, 2021.

[17] E. Kharlova, D. May, P. Musilek. Forecasting Photovoltaic Power Production using a Deep Learning Sequence to Sequence Model with Attention, 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7.

[18] A. Betti, M. Tucci, E. Crisostomi, A. Piazzi, S. Barmada, D. Thomopulos, Fault Prediction and Early-Detection in Large PV Power Plants Based on Self-Organising Maps, vol. 21(5), 1687, 2021.

[19] V. Marinakis, H. Doukas, An Advanced IoT-based System for Intelligent Energy Management in Buildings, Sensors, vol. 18(2), 610, 2018.

[20] V. Marinakis, H. Doukas, C. Karakosta, J. Psarras, An integrated system for buildings' energy-efficient automation: Application in the tertiary sector, Applied Energy, vol. 101, pp. 6-14, 2013.

[21] Weather Underground, Available at: https://www.wunderground.com (accessed 14/4/2021).

[22] Copernicus Atmosphere Data Store, Available at: https://ads.atmosphere.copernicus.eu/cdsapp#!/home (accessed 5/6/2021).

[23] Copernicus CAMS solar radiation time-series, Available at: https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-solar-radiation-timeseries?tab=overview (accessed 5/6/2021).

[24] Copernicus, Regular Validation Report - Issue #32 - S-O-N 2020, Available at: https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/EQC-solar/CAMS72_2018SC2_D72.1.3.1-2021Q2_RAD_validation_report_SON2020_v1.pdf (accessed 5/6/2021).