Applied Energy 313 (2022) 118752

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Electricity price forecasting on the day-ahead market using machine learning

Léonard Tschora [a,b], Erwan Pierre [a], Marc Plantevit [c], Céline Robardet [b,*]

[a] BCM Energy, FR-69006, Lyon, France
[b] Univ. Lyon, INSA Lyon, CNRS, LIRIS UMR5205, FR-69621, France
[c] EPITA Research and Development Laboratory (LRDE), FR-94276 Le Kremlin-Bicêtre, France

ARTICLE INFO

Keywords:
Electricity price forecasting
Machine learning
Forecast evaluation
Open-access benchmark
Explainable AI (XAI)

ABSTRACT

The price of electricity on the European market is very volatile. This is due both to its mode of production by different sources, each with its own constraints (volume of production, dependence on the weather, or production inertia), and by the difficulty of its storage. Being able to predict the prices of the next day is an important issue, to allow the development of intelligent uses of electricity. In this article, we investigate the capabilities of different machine learning techniques to accurately predict electricity prices. Specifically, we extend current state-of-the-art approaches by considering previously unused predictive features such as price histories of neighboring countries. We show that these features significantly improve the quality of forecasts, even in the current period when sudden changes are occurring. We also develop an analysis of the contribution of the different features in model prediction using Shap values, in order to shed light on how models make their prediction and to build user confidence in models.

1. Introduction

The problem of Electricity Price Forecasting (EPF) is becoming more and more challenging to solve. The applications made possible by a price forecasting model are crucial for achieving the energy transition. They allow owners of renewable energy production means to make profit on the market by anticipating price movements and promote smart applications such as self-consumption [1] or car batteries optimization [2].

At the same time, there are numerous factors that need to be taken into account to understand electricity prices. For example, energy transition policies increase the proportion of renewable energy in total production [3] and introduce new market regulations such as taxation of carbon dioxide emissions. Moreover, cross countries interconnections are multiplying and some markets such as the EPEX SPOT[1] set prices for all European countries, bringing the forecasting task to the scale of the continent.

Additionally, the pricing algorithms [4] used to balance generation and consumption can lead to price spikes, both negative and positive. These spikes can result in huge losses for unwary business owners and are difficult to handle by traditional forecasting models. Particularly, the current period is marked by repeated lockdowns that cause severe changes in the European market. The economic recovery following the COVID pandemic [5,6] also causes prices to reach up to five times the usual season price, with an increased volatility, as shown in Fig. 1.

Meanwhile, Machine Learning (ML) models are increasingly effective in solving difficult problems [7,8] and can represent complex situations [9,10]. However, they are sometimes hard to reproduce, if the described methodology and parameters are not thoroughly reported. ML models are also known to lack explainability, be difficult to interpret and are often thought of as black box models. Data analysts generally decide whether to use them or not based on a single metric evaluated solely on one dataset.

Overall, the interest of researchers and business owners in EPF is growing [11,12]. Each EPF publication proposes innovative and efficient methods, but the abundance of considered markets, forecasting tasks, time periods, models and methodologies make it difficult to compare the literature [11]. Also, it can be tricky to reproduce the results of a given article because details are often omitted and a simple lack of seeds can prevent the reproducibility of stochastic processes. Another limitation is the lack of benchmarks for model comparison, which is a gap to be filled in regards to state-of-the-art research papers for other ML applications [13,14]. Finally, the users of these models need explanations to know on which phenomena the model is based to make its predictions. This makes it possible to follow or not a surprising prediction in a very volatile market like that of electricity. We believe

* Correspondence to: Univ. Lyon, INSA Lyon, Bât Blaise Pascal, 20 avenue Albert Einstein, 69621 Villeurbanne, France.
E-mail addresses: leonard.tschora@insa-lyon.fr (L. Tschora), erwan.pierre@bcmenergy.fr (E. Pierre), marc.plantevit@epita.fr (M. Plantevit), celine.robardet@insa-lyon.fr (C. Robardet).

[1] https://www.epexspot.com/en/market-data

https://doi.org/10.1016/j.apenergy.2022.118752
Received 10 December 2021; Received in revised form 4 February 2022; Accepted 12 February 2022
Available online 4 March 2022

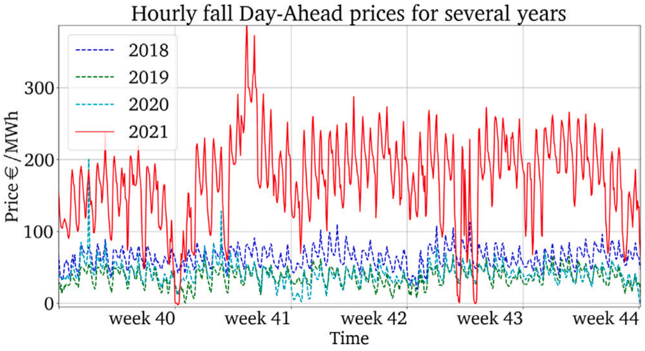

Fig. 1. Hourly Day-Ahead prices of October for the years 2018 to 2021. The prices for this period in 2021 is displayed in red and show abnormally high prices and increased volatility.

that explaining the output made by any EPF model is very important as it would help to understand what is actually captured by one model and not by another. This also helps to know which features are important in the prediction. Explainable artificial intelligence (xAI) is attracting widespread interest due to the remarkable performance of blackbox models and their need of explanations [15–18]. Although xAI has been used a lot in real applications [19], it has not yet meet the EPF problem. This article aims to provide answers to these needs. We detail below our contributions which we hope will help EPF field of research to grow richer.

1.1. Contributions

Following the guidelines introduced in a recent publication [11], we apply a rigorous, transparent and reproducible methodology for using ML models for EPF. We evaluate our ML models over three different areas of Europe on two separated test periods. We detail every step of our methodology with care and provide readers with the scripts we designed and used to make replicating our results effortless.[2] We show that the ML models are capable of correctly forecasting recent electricity prices.

We also provide EPF users with information on how the market works in addition to a reliable forecasting model. This consists in using explainable Machine Learning methods to link ML models results to real business applications by conducting a feature analysis based on Shap values [20]. Using these tools, we show the importance of using external features such as Swiss and gas prices in the EPF problem.

1.2. Paper structure

Section 2 presents related work on EPF. We state the problem that interests us and conduct a brief literature review on predictive methods used in EPF and introduce explainable artificial intelligence (xAI) main methods. We present the technical requirements on ML models, as well as their evaluation in Section 3. We also give details about Shap Values, the explainability approach used to analyze blackbox ML models. We detail the specificity of our datasets in Section 4. Due to feature availability, we differentiate the features used in the two time periods considered in our experiments. Section 5 reports our results. We first analyze the quality of the models using metrics assessing the adequacy of the predictions with true values. Then, we give to the readers the results of the explanations of the predictions, by showing the importance of the features for the model and their temporal changes in regards to exogenous events that strongly influence the volatility of the price of electricity. Finally, we conclude the paper and give directions to future work in Section 6.

2. Electricity price forecasting

Electricity markets are subject to several constraints induced by the inherent nature of this energy which requires consumption and production to be permanently matched on a continental scale. To tackle this problem, markets use pricing algorithms. For European exchanges, the EUPHEMIA [4] algorithm maximizes social welfare by solving a mixed-integer quadratic programming optimization problem. Social welfare is defined as the sum of consumer surplus, supplier surplus and cross-border trade congestion rents. EUPHEMIA ensures the highest price for producers, the lowest price for suppliers and a constant energy balance by setting Day-Ahead prices, i.e. 24-hour prices for the next day. Each market participant can submit orders until midnight for the following day. EUPHEMIA calculates the prices for each country and each hour so that they are advantageous for everyone. In doing so, it also computes cross-border flows. The Peak-Load period is defined by the EPEX exchange as the period of the day between 8:00 a.m. and 8:00 p.m., characterized by high demand. This period is used by production plant owners to issue specific orders for the entire period.

The day-ahead market forecasting problem consists of predicting, before noon, the 24-hour hourly prices for the next day. Due to the abundance of markets, business applications, and real-world forecasting methods, the EPF literature contains many innovative contributions to this problem. Several markets such as the Australian [21] and New York [22] markets are investigated, but the most studied markets are the European markets. Among them, the Spanish [23–25], the French [26], the German [27] and the Dutch [28] markets are the most studied. Note that some authors also evaluate the same models in multiple markets [29–32]. Others like [29,33] focus on market integration by including prices from neighboring areas in their prediction model.

Traditionally, auto-regressive methods were used for EPF [23,34–38]. But, over the past decade, the use of machine learning models for EPF has increased. Many models have been studied, such as Support Vector Machines (SVM) [21,27,35,39], Random Forests (RF) [22,25], Artificial Neural Networks (ANN) [24,40–42], Recurrent Neural Networks (RNN) [43] or Convolutional Neural Networks (CNN) [44]. The authors use these models to predict prices directly, but sometimes they use a more sophisticated prediction framework. For example, [45–47] predict aggregate curves (output from EUPHEMIA) and retrieves prices by interpolation. In doing so, they model the order books of market participants and mimic the real pricing mechanism. In addition, all papers published for the forecasting challenge `Gefcom` 2014 [48–51] produce probabilistic forecasts by performing quantile regression and evaluate them using the pinball loss function. Finally, [52] focuses on data augmentation and generates its datasets using autoencoders. Readers can refer to the surveys in [12,53,54] for a more comprehensive review of auto-regression and machine learning methods used for EPF. Additionally, [11] introduces several key steps to guide EPF research. Since the authors share their datasets, models, and methodology, we were convinced to follow the proposed guidelines to some extent. As part of our work, we therefore use the EPFTOOLBOX[3] which provides two models – LEAR (auto-regressive) and DNN – that we use for comparison purposes.

Unquestionably, machine learning models for EPF have become more and more accurate. On the other hand, they have become more opaque, functioning as black boxes, which limits their benefit for stakeholders [55]. Thus, explainable artificial intelligence (xAI) is an important and timely challenge in machine learning. As a consequence, the research field of xAI has grown rapidly [15–18]. In practice, the most widely used explanation methods are SHAP [56] and LIME [57] which are model-agnostic. LIME use local surrogate models to explain model output. SHAP is based on the game theoretically optimal Shapley values [58]. It explains the prediction of an instance by computing the

[2] See repository https://github.com/Leonardbcm/EPFDAML

[3] https://github.com/jeslago/epftoolbox

contribution of each feature to its associated prediction. It has been showed that LIME is related to SHAP [56]. Surprisingly, xAI has not been considered for EPF yet. This paper is the first attempt to combine both accurate model and explainable decision for the EPF problem.

3. Machine learning for EPF

Machine Learning (ML) is a branch of computer science proposing forecasting models by implementing efficient learning from data algorithms. This field has received a lot of attention in the past decades due to the abundance of available data and the growing computing power of machines. In the field of forecasting, ML models have been able to solve very complex problems in image processing [7,59,60] but also in multivariate Time Series regression [8,13,14,61–63]. As we believe that the capabilities of ML models have not yet been fully unraveled in the field of EPF, we focus on these approaches. In particular we consider four different models exposed below. We also present the metrics and tests used to compare them and the way we preprocess data. We also explain how we fix hyper-parameter values and present the recalibration strategy used to adapt models to recent changes in the data. Finally, we describe the SHAP method that we employ in our analysis to assess the importance of features in the prediction process.

3.1. Machine learning models

Support vector regressor. Support Vector Machines [62,64] are a category of models with a good mathematical background based on an optimization problem. With the use of kernels [65,66], they can be applied on complex data structures and model non-linearity. Originally designed to solve univariate forecasting problems, we adapt them to the multivariate case in two ways: 1) The ChainSVR method that uses the first forecast to predict the second one, the second forecast to predict the third one, and so on; 2) The MultiSVR that uses one model per time series in Y, so n_o in total. We use the method SVR as implemented in `scikit-learn` [67].

Random forest regressor. Random Forest Regressor models (RFR) are widely used ML models both in the field of EPF [22,25,30] and in forecasting tasks in general [13,14]. They consist of a combination of a several Decision Tree Regressor (DTR) that are trained using different subsets of the data. The Bagging [68] method used in this paper outputs the average of their predictions. We use `scikit-learn`'s implementation [67] of RFR.

Deep neural networks. The model capabilities and tremendous range of application made Deep Neural Networks (DNN) the center of interest of numerous researchers in EPF [11,29,30,41] but also in forecasting tasks in general [69–71]. The DNNs we use have $\ell + 2$ layers stacked sequentially. The number of neurons of the first and the last layer are respectively n_c and n_o, the second dimensions of X and Y respectively, the other layers having $(n_1, \dots, n_\ell)$ neurons. These hyper-parameters $(\ell, n_1, \dots, n_\ell)$ are set with a grid search. The model is trained using a gradient descent algorithm of the forecast errors back to the network weights.

Convolutional neural network. Convolutional Neural Networks (CNN) are a variant of Deep Neural Networks which became popular for their image processing capabilities [7,59,72,73]. They are now also used for multivariate time series regression tasks [8,9,74,75] and in particular, for EPF [30,44,76]. The eponymous convolutional layers combined with pooling layers are the particularity of CNNs. By applying numerous filters on the data, convolutional layers extract complex patterns that are then generalized by a pooling operation to provide complex feature representations of the input. We use the keras[4] implementation with tensorflow [77] backend to implement our Neural Networks models (DNNs and CNNs).

In order to ensure the consistency of the results obtained with those of [11] (DNN models) and to compare ML models with auto-regressive ones (LEAR models), we also reproduce their results for four LEAR models and four DNN models. The LEAR models are denoted LEAR$_{52}$, LEAR$_{84}$, LEAR$_{1092}$ and LEAR$_{1456}$, in reference to their respective calibration window size. The DNN models are denoted DNN$_1$, DNN$_2$, DNN$_3$, DNN$_4$.

3.2. Evaluation metrics and test

Let Y_d^h be the price for a day d and an hour h of a given country, and let $\hat{Y}_d^h$ be the values predicted by a model. The comparison of these values is used to evaluate and test the quality of a model, but also to learn it, through the loss function used to adjust the parameters of the model.

Metrics. The most commonly used metric to evaluate the quality of a model in the field of EPF is the Mean Absolute Error (MAE):

$$MAE(Y, \hat{Y}) = \frac{1}{n_d} \sum_{d=1}^{n_d} \sum_{h=1}^{n_0} |Y_d^h - \hat{Y}_d^h|$$

It allows business owners to quickly estimate how they could use a forecasting model to generate profit. However, since electricity prices can range from −500 to 3000 €/MWh in the European markets, it is useful to use a relative error measure. While the Mean Absolute Percentage Error (MAPE) is usually used for this purpose, we prefer employ the Symmetric Mean Absolute Percentage Error (SMAPE). Indeed, prices close to 0 that are incorrectly predicted lead to a unnecessary high MAPE, which is not the case with SMAPE values:

$$MAPE(Y, \hat{Y}) = \frac{100}{n_d} \sum_{d=1}^{n_d} \sum_{h=1}^{n_o} \frac{|Y_d^h - \hat{Y}_d^h|}{|Y_d^h|}$$

$$SMAPE(Y, \hat{Y}) = \frac{100}{n_d} \sum_{d=1}^{n_d} \sum_{h=1}^{n_o} \frac{|Y_d^h - \hat{Y}_d^h|}{\frac{1}{2}\left(|Y_d^h| + |\hat{Y}_d^h|\right)}$$

We also consider a new metric called the Daily Average Error (DAE). It consists in computing the MAE between the average predicted price for a day and the real average price. This metric is very useful for trading-related activities, when one speculates on the average price for a given day.

$$DAE(Y, \hat{Y}) = \frac{1}{n_d} \sum_{d=1}^{n_d} \left| \frac{1}{n_o} \sum_{h=1}^{n_o} y_d^h - \frac{1}{n_o} \sum_{h=1}^{n_o} \hat{y}_d^h \right|$$

Next, to enable cross-dataset comparison, we use the Relative Mean Absolute Error. The idea is to compare the MAE of a model with the MAE of a naive forecaster. As naive forecaster, we use the following strategy:

$$\hat{Y}_{d,naive}^h = \begin{cases} Y_{d-1}^h & \text{if } d \text{ is a week day} \\ Y_{d-7}^h & \text{otherwise} \end{cases}$$

$$RMAE(Y, \hat{Y}, \hat{Y}_{naive}) = \frac{MAE(Y, \hat{Y})}{MAE(Y, \hat{Y}_{naive})}$$

Diebold & mariano test. We use the Diebold & Mariano [78,79] test to perform more robust model comparison. Instead of averaging a loss g across the entire dataset, it computes the loss difference d between two model predictions Y_1 and Y_2. A one sided z-test is then performed to assess if the second model forecasts are significantly better than the first ones :

$$\begin{aligned} d(Y, \hat{Y}_1, \hat{Y}_2) &= g(Y - \hat{Y}_1) - g(Y - \hat{Y}_2) \\ H_0 &: E(d(Y, \hat{Y}_1, \hat{Y}_2)) > 0 \\ H_1 &: E(d(Y, \hat{Y}_1, \hat{Y}_2)) \leq 0 \end{aligned}$$

If the obtained p-value is lower than a fixed threshold of 0.05, then H_0 is rejected and we can conclude that the first model is better than the second one.

We use the absolute loss $g(Y, \hat{Y}) = MAE(Y, \hat{Y})$ in our experiments as it better reflects business applications.

[4] https://keras.io

Loss. We use the LogCosH loss function for training Neural Networks models (DNNs and CNNs). It combines the benefits of both MAE and Mean Squared Error by being approximately equivalent to $\frac{(Y-\hat{Y})}{2}$ when $Y-\hat{Y}$ is small, and to $|Y-\hat{Y}|-\log(2)$ when differences are large. Due to the presence of spikes in electricity prices, it is useful not to put too much weight on outliers:

$$LogCosH(Y,\hat{Y}) = \log(\frac{e^{Y-\hat{Y}}+e^{\hat{Y}-Y}}{2})$$

3.3. Data preprocessing

Data scaling is critical during ML model training. Most algorithms require that both the input (X) and output (Y) data are pre-processed. To this aim, we design simple data pipelines to process the features and target variables of our datasets. We distinguish the scaler used to process the input data X from the transformer used for processing the predicted values Y. We consider these two functions as hyper-parameters with four different possibilities for each of them: (1) the standard scaler that standardizes data so it has a 0 mean and 1 variance, (2) the median scaler, a outlier-robust version of standard scaler using the median and median average deviation, and (3) their combination with the $arcsinh$ function [80] or not:

$$SS(X) = \frac{X-\mu_x}{\sigma_x^2} \tag{1}$$

$$MS(X) = \frac{X-median_x}{MAD_x^2} \tag{2}$$

$$arcsinh(X,f) = \log\left(f(X)+\sqrt{f(X)^2+1}\right), \tag{3}$$

with f either SS or MS.

3.4. Hyper-parameters search

Despite their high modeling power, ML models suffer from a critical issue that is hyper-parameter optimization. Hyper-parameters must be configured before training the model on the data. They need to be tuned for optimal results. This is done by testing numerous combinations of hyper-parameters and selecting the optimal one. As this part is very time consuming, we use a Randomized Grid Search [81] that samples 4000 hyper-parameter combinations for each models in a pre-defined search space. Details of the search spaces for each model are available on our repository.[5]

3.5. Recalibration

Another drawback of ML models is their implicit assumption that the future will be similar to the past. However, as seen in Fig. 1, electricity prices can be very volatile and sudden unpredictable changes can drastically modify the prices, such as the Covid lockdown [5] or the European energy gaz crisis of fall 2021 [82]. Those changes are critical, for example, [29]'s model gets confusing results while forecasting Belgian prices due a sudden change in the generation patterns. To leverage such problems, [22] uses an online Random Forest method to keep the forecasting model up to date, [52] generate more current data using autoencoders and [11] uses model recalibration. Recalibration consists in retraining the model with most recent data, that is to say using $X_1,\dots,X_{i-1}$ and $Y_1,\dots,Y_{i-1}$ to train the model before forecasting a new sample X_i, $(X_1,Y_1),\dots,(X_{i-1},Y_{i-1})$ being in the test set. However, computational costs are induced by this method as the models have to be re-trained from scratch for each new sample to predict. Each evaluation step requires as many model trainings as there are samples in the test set. The search of optimal hyper-parameters, that is based on the evaluation of numerous combinations, becomes too costly. We decided to evaluate the performance of a combination on the basic forecasts, without recalibration.

3.6. Shapley and SHAP values

While the features all together contribute to the prediction process, it is difficult to measure the importance of each of them in the decision. Indeed, there are many correlations between the variables, and properly measuring the impact of each variable requires taking the interactions into account. Shapley values were defined within the framework of game theory in order to fairly distribute a gain among several players in a cooperative game. Fair means that the contribution of the players is taken into account in obtaining the gain. This means that a player is not only paid for what he is able to gain when he is alone, but also for his contribution to the group when interacting with other players. To calculate the Shapley value associated with the feature i, ϕ_i, it is necessary to calculate for each coalition Z in which i does not appear, the difference in gain $f(Z\cup\{i\})-f(Z)$. This makes it possible to compare the gain obtained from the coalition with and without i, in order to measure its impact when it collaborates with the set Z of features. If this difference is positive, it means that feature i contributes positively to this coalition. Conversely, if the difference is negative, it means that i penalizes the group. Finally, if the difference is zero, this indicates that i does not contribute anything to this group. The gain to be distributed is here the difference between the forecast and the average of the forecasts.

To specify more formally Shapley values, it is necessary to define a mapping $h_x(Z)$ that maps the input vector x to the same vector where features that are not in Z are missing. We also define $f_x(Z) = E[f(x) \mid x_Z]$ the expected value of f conditioned on a subset Z of the input features. The Shapley values are a weighted average of all possible differences between the coalitions of features including and not including i:

$$\phi_i(x) = \sum_{Z\subseteq F\setminus\{i\}} \frac{|Z|!(|F|-|Z|-1)!}{|F|!}\left(f_x(Z\cup\{i\})-f_x(Z)\right)$$

where F is the set of all input features.

The calculation of a Shapley coefficient poses two difficulties: estimating the conditional expectations and dealing with the combinatorial explosion of the number of coalitions to go through, when the number n_c of features increases. The number of coalitions to be covered is exponential, in 2^{n_c}. [56] introduces the concept of Shapley kernel to approximate Shapley values and makes it possible the use of this approach on real-world dataset such as EPF ones. We use python's SHAP[6] package to compute the SHAP values of our models, using a total of 2500 subsets per forecasts.

The method SHAP (SHapley Additive exPlanations) uses the Shapley values to compute an additive explanatory model g that is a linear combination of Shapley values:

$$g(x') = \phi_0 + \sum_{i=1}^{n_c}\phi_i x'_i$$

with ϕ_0 the average output of the model, ϕ_i the explained effect of feature i and x' a binary encoding of instance x. This explanatory model is constrained to be roughly equal to f in the vicinity of x.

4. Datasets

Many multivariate time series forecasting research articles [13,14] recommend to evaluate models on several datasets as the behavior of a same algorithm can be very different depending on unknown characteristics of the dataset. The relative performances of several models can even vary and considering a large number of datasets makes it possible to have a more robust evaluation of the model performances. To assess the specific qualities of a model, it is therefore relevant to consider datasets from different countries. Indeed, the energy mixes are

[5] See repository https://github.com/Leonardbcm/EPFDAML

[6] https://shap.readthedocs.io/en/latest/index.html

Table 1
Exogenous inputs of EPFTOOLBOX dataset. Each dataset is composed of the Day-Ahead prices for the specified country and 2 exogenous features.

Dataset	Exogenous input 1	Exogenous input 2
FR	Consumption forecast	Production forecast
BE	French consumption forecast	French production forecast
DE	Amprion consumption forecast	Amprion, TenneT, 50 H renewable forecasts

very different from one country to another and have a strong influence on the dynamics of the prices of electricity.

To build predictive models of electricity prices, we extend the classically considered datasets [11], called hereafter SOTA, by adding new attributes as predictive features and considering more recent data. These datasets and their specificity are presented below.

4.1. SOTA datasets

We consider three datasets from [11]. These datasets contain electricity prices for 6 years for three geographical areas: France (FR), Germany (DE), and Belgium (BE). Each dataset includes next day prices and has two additional exogenous features given in Table 1.

Electricity price datasets are a multivariate time series made of daily data. Those datasets can be reconfigured into a (X, Y) couple suitable to learn machine learning models. The predictive data is represented by a two dimensional matrix $X \in \mathbb{R}^{n_d \times n_c}$ whose rows represent days and columns are n_c predictive time-dependent values. The values to be predicted correspond to another matrix $Y \in \mathbb{R}^{n_d \times n_o}$, whose rows also stand for the days and columns are the n_o day-ahead prices to be predicted: $Y_d = \left(Y_{d+1}^1, \dots, Y_{d+1}^{n_0}\right)$. To model the time series aspect of the features, X includes the prices of the current day, those of the day before, two days before and the previous week (1, 2, 3 and 7 days lag). Exogenous features are included for the day, the day before and the previous week. In addition to these 240 characteristics, the day of the week is also encoded as an integer and added to the matrix X. Indeed, electricity prices are non-stationary time series and exhibit seasonal trends captured by this additional feature. All features (prices and exogenous) are provided with hourly granularity. Thus, the predictive matrix X is as follows:

$$X_d = \big(Y_{d-1}, Y_{d-2}, Y_{d-3}, Y_{d-7}, E1_d, E1_{d-1}, E1_{d-7}, \\ E2_d, E2_{d-1}, E2_{d-7}, \text{DayOfWeek}\big) \text{ with } n_c = 241.$$

In order to forecast 24-hour prices for the next day, the datasets are reshaped so that for one day d, Y_d contains all 24 prices for the next day: $Y_d = \left(Y_{d+1}^1, \dots, Y_{d+1}^{24}\right)$.

4.2. Enriched datasets

For the enriched datasets considered in this study, we focus on three European countries: France, Germany and Belgium. These countries are at the same time geographically close, but have features that make them unique. For example, the French generation fleet is 75% composed of nuclear power plants [3] which are to some extent controllable, unlike wind turbines which constitute 45% of the German generation system. As a result, prices in Germany tend to be more volatile and sometimes reach negative values. In addition, French consumption is mainly heat-sensitive due to the massive use of electric heaters leading to higher prices in the winter period. Belgium, for its part, has a much lower level of consumption and can be used to transport energy from France to Germany or the Netherlands.

From these data we build four datasets, three (FR, DE, BE) comprising the data of each country taken individually, and a fourth (Multi-Output) merging together the data of the three countries. With this dataset, we seek to forecast the prices of the three countries at the same time. Due to the pricing algorithm, all European prices are set at the same time and we want to model this phenomenon.

Electricity day-ahead price is fixed by EUPHEMIA through the coupling of different markets where energy transactions can involve sellers and buyers from different countries, only limited by the constraints of the electricity network. All bilateral interconnections make it possible to transport less expensive production assets from one country to another with an important demand. Thus, the price within a country is highly dependent on exogenous factors in surrounding countries. This is why we have included production and consumption forecasts from neighboring countries in our datasets. Similarly, we used Dutch, Spanish and Swiss prices. Swiss prices are attractive as they are available every day at 11.15 am and can be used in a forecasting model before the European market closes at noon.

Another aspect that can strongly influence the prediction are the dates, especially the days of the week that involve differentiated human activity and therefore impact energy consumption and production. But, as shown by [83], the seasonality of the electricity market is not only dependent on the day of the week. We therefore propose to incorporate various date dummy variables into our enriched dataset. We decided to include weekday, week number, day of month and month number as predictive functions. To better integrate these cyclic data into our ML models, we apply a circular encoding transformation f of a cyclic feature that encodes the original feature of the domain value C (with cardinality α) into two numeric values:

$$f : C \mapsto \mathbb{R}^2 \\ x \mapsto (\sin(\frac{2\pi x}{\alpha}), \cos(\frac{2\pi x}{\alpha}))$$

Finally, we also integrate gas prices. Indeed, to maximize social welfare, the EUPHEMIA algorithm favors the power plant with the lowest marginal cost. Accordingly, there is an order of merit for the technology of production plants. Gas-fired power plants are one of the cheapest ways of generating electricity among other coal or oil-fired thermal power plants. However, its marginal cost is a function of gas prices. Therefore, depending on the country's energy mix, gas prices are an important feature of electricity prices. We therefore decided to include the EGSI gas index[7] in our dataset. As this index is available every day at 6pm, after the market closure, it has to be included for predicting prices 2 days after.

As previously, to model the time series aspect of these features, X contains the country's prices for the previous day, those of two days before, three days before and the previous week (1, 2, 3 and 7 days lag). Other features (see Table 2) are included for the day before, the previous week, and if possible the current day. Indeed, production and consumption forecasts as well as Swiss prices are available for the day to be forecast before noon. Then, with the exception of gas prices and date dummies, all features are included for the 24 h of the day. The datasets therefore have $n_c = 24 \times n_f^l \times n_f + 8 + 1$ columns, with n_f the number of features as described in Table 2 and n_f^l the number of shift days for a given feature f.

4.3. Train/test splits

In [11], the authors provide open-access benchmark datasets for the 6-year period between 2011 and 2016. A good practice in the field of machine learning is to evaluate models over the same time period to allow comparison of results. We therefore start our analysis by evaluating our models on the same data (see dataset description T_1 in Table 3).

It is also important to extend our study to the current period, whose peculiarities are a source of evaluation of the robustness and adaptability of the models in a context of high variability. Therefore, we

[7] https://www.boursorama.com/cours/1rPGTT/

Table 2
Composition of the datasets for each country and the two time periods.

Features	FR		DE		BE	
	T_1	T_2	T_1	T_2	T_1	T_2
French prices	Target	Target	✓	✓	✓	✓
German prices	✓	✓	Target	Target	✓	✓
Belgian prices	✓	✓	✓	✓	Target	Target
Dutch prices	✓	✓	✓	✓	✓	✓
Spanish prices	✓	✓				
Swiss prices	✓	✓	✓	✓		
French consumption forecast	✓	✓	✓	✓	✓	✓
German consumption forecast	✓	✓	✓	✓	✓	✓
Belgian consumption forecast		✓		✓		✓
French production forecast	✓	✓	✓	✓	✓	✓
German renewable energy forecast	✓	✓	✓	✓	✓	✓
Belgian renewable energy forecast		✓		✓		✓
French gas prices		✓		✓		✓
Date dummies	✓	✓	✓	✓	✓	✓

Table 3
Time period of data used for training (learning of model parameters), validation (determining hyperparameter values), and testing (evaluating models) for EPFTOOLBOX datasets. As we use the first seven days of the dataset as the input features, the train dataset starts seven days after the first data sample.

Period	Train start	Validation start	Test start	Test end
T_1	2011–01–16	2014-01-07	2015–01–04	2016–12–31
T_2	2016–01–01	2019-01-01	2020–01–01	2021-12-31

consider a second dataset with 3 years from 2016 to 2019 for training and two test years from 2020 to 2021 (see dataset description T_2 in Table 3). Recent electricity prices present a more difficult challenge for prediction because the lock down related to Covid-19 has caused massive changes in the European market. Furthermore, 2021 is marked by a limited energy crisis. In addition, since 2015, the ENTSOE transparency platform[8] has brought together and published data from almost all European TSOs in free access. This results in much more available data. The features of our datasets therefore vary depending on the period considered as described in Table 2.

5. Evaluation of the models on the different datasets

The objective of this section is to evaluate the different models of machine learning. First, we measure the impact of considering the additional features on the accuracy of predictions. We also evaluate the interest of simultaneously predicting the price of electricity in several countries. Then, we propose to study the models from an XAI point of view, to identify on which variables the predictions are based.

5.1. How well do the models performs?

We present the performance measures of the different models in Table 4. We both compare the models to each other, but also evaluate the impact of adding features on the predictions. To do a fair comparison with [11], we consider the T_1 time period that was used in this paper. For a better interpretability of the multi-country models, we display the metrics for each of the forecasts of the three countries.

First, we can observe that using additional features to predict prices always increases performance. Each model gets better metric values with the use of the new features with up to 15% gain. We support this finding by highlighting the p-values of Diebold & Mariano tests between models trained on SOTA datasets and their counterparts trained on enriched datasets in Table 5 (A). We can observe that this difference is statistically significant for the vast majority of countries and models (values in bold). These tables also reveal that Belgian prices are more difficult to predict than other country prices in the single-country framework. The RMAE in Table 4 indicates that the best model for that country only achieves a fraction of 0.7 of the error of a naive forecaster. The other datasets have an RMAE lower than 0.6, or even 0.45 on German dataset. We believe this is due to the fact that Belgian consumption and production forecasts are not available for this period. This is discussed in more detail in Section 5.2. Another conclusion from these experiments is that Random Forest models do not predict prices accurately. Their metric values are always significantly higher than those of the other models on all datasets. The enriched datasets still increase performances but they do not necessarily outperform other models based on SOTA datasets. It also appears that CNN models are not state-of-the-art forecasting models for EPF. Even though they obtain reasonable metric values on the enriched datasets, they never significantly outperform the DNN or SVR models and this for all datasets. We believe that the data provided to CNN models is not suitable for convolutions. CNN models are tailored for extracting meaningful patterns among raw features, such as basic geometric shapes on an image. We feed it with data such as production or generation forecasts which is a high-level representation of meteorological data. Moreover, we reshape our data as 32 × 24, which is a very small amount of data compared to SOTA CNN models. For example, the AlexNet model introduced in [7] works on images of 224 × 224 pixels. Finally, still considering Table 4 and the p-values in Table 5 (B) and (C), we see that the interest of jointly predicting the prices of several countries is mixed. The multi-country forecast model reduces forecast quality by up to 5%. This reduction is significant on 4 of the 5 models in France and Germany (column C). However, it significantly increases the performance of 3 out of 5 models in Belgium (column B). Merging the three datasets did not add any crucial and previously unknown information to the French and German datasets. On the other hand, it allows the model to use Swiss prices to predict Belgian prices. We believe this explains the significant increase in Belgium's performance.

We now study the robustness of these observations by considering the time period T_2. We present the metric values obtained for this period in Table 6. We can make the following observations. First, the best absolute metrics (MAE & DAE) increased by almost a factor of two over the T_2 period. This is not surprising as price levels also shift from 38€44/MWh on average in 2015 to 109€11/MWh in 2021 in France. However, the RMAE decreased from 0.55 to 0.46 for the French MultiSVR, from 0.44 to 0.42 for the German DNN and from 0.67 to 0.57 for the Belgian MultiSVR, which shows that the models are performing better against the baseline than for the previous period. Our ML models successfully integrated sudden changes in electricity markets. Second, it appears that the Belgian dataset experiences the most significant performance increase. The availability of Belgian consumption and production forecasts made this data set easier to predict than for the previous period. However, it is still the most difficult country to predict because we do not use Swiss prices, as this country does not border

[8] https://transparency.entsoe.eu

Table 4
Performance metrics over the period T_1. The multi-output models' metrics are reported country by country. Best performance metrics are always obtained on enriched datasets and for Belgium on the multi-output models.

	Metric	SOTA datasets													Enriched datasets					Multi-output models				
		LEAR				DNN				CNN	DNN	RF	SVR		CNN	DNN	RF	SVR		CNN	DNN	RF	SVR	
		56	84	1092	1456	1	2	3	4				Chain	Multi				Chain	Multi				Chain	Multi
FR	smape	13.32	13.41	13.57	14.59	12.00	11.65	11.75	11.51	12.05	11.57	13.42	11.23	11.26	10.80	11.12	11.81	**10.43**	10.56	11.07	10.95	12.44	10.66	10.67
	mae	4.63	4.58	4.35	4.48	4.34	4.15	4.17	4.12	4.27	4.15	4.73	4.03	4.06	3.79	3.89	4.11	**3.65**	3.67	3.92	3.85	4.30	3.71	3.71
	dae	3.37	3.34	3.25	3.38	3.35	3.13	3.09	3.09	3.14	3.05	3.46	3.02	3.02	2.67	2.71	2.90	2.57	**2.56**	2.81	2.78	2.99	2.61	2.57
	rmae	0.69	0.68	0.65	0.67	0.65	0.62	0.62	0.61	0.64	0.62	0.71	0.60	0.61	0.57	0.58	0.61	**0.54**	0.55	0.59	0.57	0.64	0.55	0.55
DE	smape	15.25	15.16	17.31	17.94	14.27	14.49	14.25	14.20	16.26	14.47	17.43	14.48	14.53	14.24	**13.56**	15.83	13.84	14.40	14.42	13.96	16.45	14.22	14.29
	mae	3.64	3.59	3.61	3.72	3.27	3.34	3.22	3.23	3.63	3.27	4.11	3.27	3.28	3.19	**3.12**	3.72	3.15	3.25	3.29	3.24	4.11	3.26	3.28
	dae	2.54	2.53	2.65	2.74	2.32	2.46	2.29	2.33	2.58	2.27	2.79	2.38	2.38	2.16	**2.12**	2.37	2.22	2.31	2.26	2.22	2.71	2.33	2.34
	rmae	0.50	0.49	0.50	0.51	0.45	0.46	0.44	0.44	0.50	0.45	0.57	0.45	0.45	0.44	**0.43**	0.51	**0.43**	0.45	0.45	0.45	0.57	0.45	0.45
BE	smape	17.02	17.32	17.20	17.75	15.77	14.79	15.77	15.17	15.20	14.59	15.37	14.50	14.46	14.12	14.82	15.22	14.35	14.28	13.8	13.60	15.50	13.50	**13.47**
	mae	7.28	7.32	6.68	6.73	6.84	6.37	6.76	6.50	6.43	6.25	6.55	6.41	6.25	6.14	6.33	6.50	6.11	6.14	6.01	5.87	6.66	**5.88**	5.90
	dae	5.18	5.20	4.84	4.91	5.15	4.67	5.00	4.75	4.62	4.50	4.79	4.77	4.62	4.51	4.69	4.85	4.58	4.54	4.28	4.20	4.98	**4.35**	4.29
	rmae	0.82	0.83	0.76	0.76	0.78	0.72	0.77	0.74	0.73	0.71	0.74	0.73	0.71	0.70	0.72	0.74	0.69	0.70	0.68	**0.67**	0.75	**0.67**	**0.67**

Table 5
P-values of the Diebold & Mariano tests for the T_1 period. (A) the test compares models trained on SOTA datasets with the same trained on enriched datasets. The null hypothesis states the enriched dataset has lower metric values than SOTA dataset models. With a threshold $\alpha = 5\%$, models in bold are significantly better when trained on the enriched datasets. (B) compares the single country forecasting models with the multi-country ones. The null hypothesis states the multi-country forecasting models are better than single-country ones on enriched datasets (values in bold). (C) The null hypothesis states the single-country forecasting models are better than multi-country ones on enriched datasets (values in bold).

Country	Model	A	B	C
		H_o : $m_{SOTA} > m_{enriched}$	H_o : $m_{enriched} > m_{multi}$	H_o : $m_{multi} > m_{enriched}$
FR	CNN	**0**	1	**0**
	DNN	**0**	0.176	0.824
	RF	**0**	1	**0**
	ChainSVR	**0**	0.989	**0.011**
	MultiSVR	**0**	0.975	**0.025**
DE	CNN	**0**	1	**0**
	DNN	**0.001**	0.999	**0.001**
	RF	**0**	1	**0**
	ChainSVR	**0.003**	1	**0**
	MultiSVR	0.219	0.949	0.501
BE	CNN	**0**	1	**0**
	DNN	0.919	**0**	1
	RF	0.117	0.845	0.155
	ChainSVR	0.998	**0**	1
	MultiSVR	0.991	**0**	1

Table 6
Performance metrics over period T_2.

Country	Metric	Enriched Datasets				
		CNN	DNN	RF	SVR	
					Chain	Multi
FR	smape	19.75	15.97	17.33	**14.23**	**14.23**
	mae	10.40	7.96	9.41	6.86	**6.61**
	dae	7.65	5.70	7.06	5.10	**4.74**
	rmae	0.73	0.56	0.66	0.48	**0.46**
DE	smape	20.36	**18.79**	22.35	18.80	19.45
	mae	8.66	**7.66**	10.77	8.44	8.85
	dae	6.53	**5.13**	7.77	6.25	6.62
	rmae	0.47	**0.42**	0.58	0.46	0.48
BE	smape	24.85	21.65	21.60	**18.93**	19.17
	mae	14.18	11.86	12.30	**9.35**	9.51
	dae	10.09	9.37	9.68	**6.67**	6.78
	rmae	0.88	0.73	0.76	**0.58**	0.59

Belgium. Third, the differences in performance between the models are greater over this period. We clearly identify that the SVR models are better on the French and Belgian datasets while the DNN is the best model on the German dataset. The DM test pvalues in Fig. 2 confirm that this difference is significant. On this figure, colored squares at coordinates (i, j) indicates that the forecasts of model i are significantly more accurate than forecasts of model j. We clearly identify green columns for the SVR models in France and Belgium, indicating that the MultiSVR and ChainSVR significantly outperform the other models. For these countries, the DNN model outperforms the RF and CNN models. Finally, the CNN model is significantly less efficient than all the other models. For Germany, only the DNN significantly outperforms all other models, while the RF model is significantly outperformed by all other models.

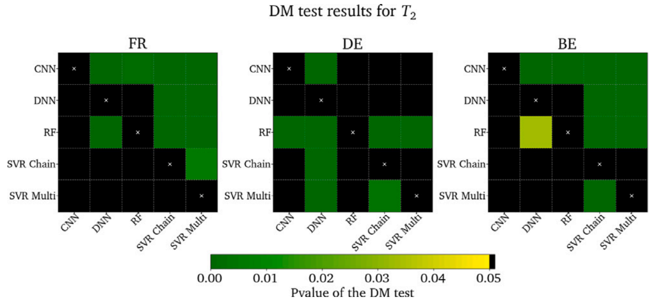

Fig. 2. P-values of the Diebold & Mariano tests computed on the recalibrated forecasts on period T_2. Colored squares in (i, j) indicates that the forecasts of model i are significantly more accurate than forecasts of model j. Green columns indicate that the corresponding models are significantly better than every other. Black lines indicate that the model on the y-axis' forecasts are significantly worse than every other.

5.2. Forecast explanations

We have seen that the performance of the models is generally of good quality. Some of these models even have equivalent performance and it is difficult to decide between them. Moreover, to increase the confidence in the predictions given by models, it is necessary to be able to explain them and to identify the most important characteristics in the decision-making process. This allows us to better appreciate their quality and better understand the phenomena involved in price prediction.

We have seen that adding features dramatically improves model performance for the vast majority of datasets and models. A legitimate question is then to ask which features contribute the most to the prediction? Different techniques exist to explain the decision process of a model [57,84]. In the following, we consider the SHAP value approach [85], a method that assigns each feature a value that reflects its importance in the prediction process. $\Phi_c^{d,o}$ designates a SHAP value and denote the contribution of a column c to the output o on day d.

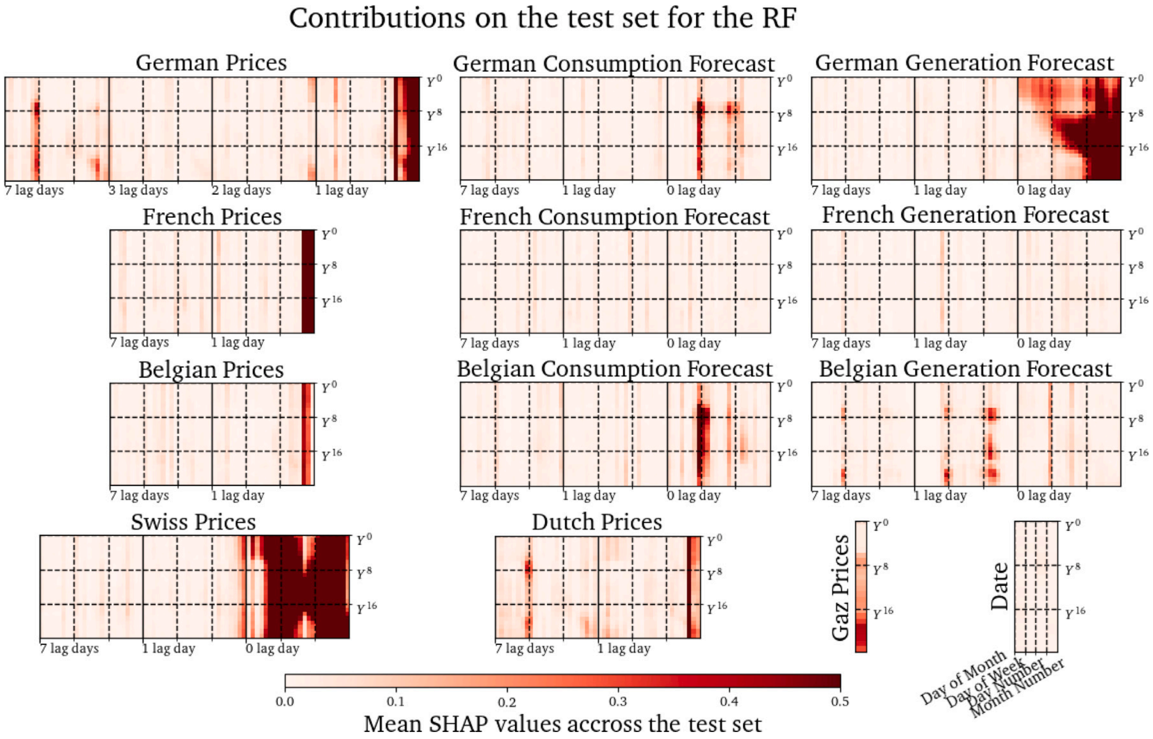

Fig. 3. Average feature contribution of the RF model for the German dataset. Each subplot displays the contributions of a single feature on all target variables. The target variables are on the y-axis while the lag days and hours are represented on the x-axis. A red square on subplot f at the coordinates $((l,h),o)$ means that the contribution of feature f with l lag days at hour h is high on average for the output o. We observe numerous contributions close to 0, meaning that some features are omitted.

Note that a column $c = (f, l, h)$ refers to the hour h of a feature f with l days lag. We also divide the contribution of each column c so that $\sum_{c=1}^{n_c} \bar{\Phi}_c^{d,o} = 1$.

We first focus on explaining the performances gaps between models. Results for period T_2 on the German dataset are presented on Figs. 3, 4, 5. Each subplot corresponds to a feature f. On the x-axis are all possible lags in hours for this feature, while the outputs are shown on the y-axis. For a feature f, we display $\bar{\Phi}^o_{f,l,h}$ at coordinates $((l,h),o)$, the average SHAP values over all days d $\frac{1}{n_d}\sum_d \Phi^{d,o}_{f,l,h}$. We also report the average contribution for each feature $\bar{\Phi}_f$ as a percentage of the total contribution in Table 7:

$$\bar{\Phi}_f = \frac{1}{n_d n_o} \sum_{d=1,o=1}^{n_d,n_o} \frac{1}{n_h n_l} \sum_{h=1,l=1}^{n_h,n_l} \Phi^{d,o}_{f,l,h}.$$

The average contribution for each day lag $\bar{\Phi}_l$ is shown in Table 8:

$$\bar{\Phi}_l = \frac{1}{n_d n_o} \sum_{d=1,o=1}^{n_d,n_o} \frac{1}{n_h n_f} \sum_{h=1,f=1}^{n_h,n_f} \Phi^{d,o}_{f,l,h}.$$

Lastly, we study the evolution of feature contribution along time. Particularly, we study the effects of the three Covid lockdowns in France on the daily mean unit contribution that we defined as

$$\bar{\Phi}^d_l = \frac{1}{n_l n_h h} \sum_{l=1,h=1}^{n_l,n_h} \frac{1}{n_o} \sum_{o=1}^{n_o} \Phi^{d,o}_{f,l,h}.$$

Fig. 6 displays these measures.

Fig. 3 presents the feature contribution of the Random Forest model for the German dataset. We can observe that most of the feature contributions are close to 0 or are used uniformly to predict all hourly prices over 24, forming a vertical line of red squares ($\bar{\Phi}^o_{f,l,h}$ is high $\forall o = 1 : n_o$). We relate this observation to the way RF models are trained: the Multi-Output Decision Tree algorithm chooses a division that satisfies the split criterion for all target variables. Therefore, at least on the higher nodes, the same characteristics are used to determine all the target features and their contributions is thus high. Moreover, we see in Table 7 that the RF models do not use all the information of the different features with the same importance: most of the contributions are made by the Swiss prices and by the country-specific prices (for French and German datasets). Finally, from Table 8, we observe that they barely use the feature with a two, three or seven day lag. We believe that these three facts explain why RF models perform significantly worse than any other model on every data set.

MultiSVR contributions, shown in Fig. 4, display diagonals of red squares that occur when $\bar{\Phi}^o_{f,l,h}$ is high $\forall o = h$. This means that a column $c = (f, l, h)$ contributes to target variable o only if $o = h$. This is most visible for the German generation forecast for the day to predict. Indeed, the generation forecast in Germany is a volatile feature (half of it comes from wind generation) that market players usually take into account for making their order books and helps estimating the prices. Patterns are hard to identify in the foreign features with lag days such as French consumption or generation forecasts, even though the contributions for these features are not null. We can observe partial diagonals in German, Dutch or Swiss price features. Due to market coupling, prices at a given hour from neighboring countries are sometimes identical, hence they constitute an important feature for prediction. Lastly, we observe strong contributions from evening prices with a 1-day lag, such as German, French and Dutch prices. The previous evening's prices are closest in time to the prices we aim to predict and are therefore an interesting feature well captured by the model.

The DNN model contributions in Fig. 5 display centered group of red squares: high $\bar{\Phi}^o_{f,l,h}$ for $h = 8$ am to 8 pm and target variables $o = 8$ am to 8 pm. For instance, the German consumption forecasts from 8 am to 8 pm highly contributes to the German prices forecasts from 8 am to 8 pm. Peak-load specific orders can be issued by market players during those hours, and this is most used by power plant owners to allow them to either turn their plant on or shutting it down during those 12 h. We also identify diagonal patterns for several features such as the Generation forecasts or German and Dutch prices. The patterns observed on this model give a finer representation of its use of input

Table 7

Summary of average contributions by feature category over time period T_2. The contributions are summed for all targets, all times and all offsets for each category. The last two columns display the weight of the characteristics of foreign countries in the total contribution.

	Model	FR consumption Forecast	FR generation Forecast	DE consumption Forecast	DE generation Forecast	BE consumption Forecast	BE production Forecast	FR price	DE price	BE price	NL price	ES price	CH price	Features from country	Foreign features	Date indicators & gaz price
FR	CNN	7.11	6.65	7.59	5.45	5.13	22.18	19.38	5.09	3.45	4.59	3.37	8.13	33.14	64.97	1.89
	DNN	7.10	5.13	4.28	5.61	5.15	7.55	15.49	5.60	3.82	4.78	5.63	28.39	27.71	70.80	1.49
	RF	1.33	1.00	1.58	1.13	1.33	2.19	**15.49**	1.21	2.30	1.03	1.06	**70.09**	17.82	81.92	0.26
	SVR Chain	4.10	9.81	2.86	4.18	4.23	6.22	17.46	6.16	4.47	6.90	5.62	26.56	31.37	67.21	1.42
	SVR Multi	4.10	9.06	2.68	4.55	4.06	6.90	17.16	5.69	4.36	6.16	5.65	27.98	30.32	68.03	1.65
DE	CNN	4.39	7.11	6.44	18.13	6.03	6.23	5.06	22.64	3.90	8.24	–	9.46	47.21	50.42	2.37
	DNN	4.11	4.98	5.08	18.10	4.78	6.89	3.88	20.95	3.67	9.27	–	16.66	44.13	54.24	1.63
	RF	0.91	0.87	1.85	11.17	2.54	1.77	3.53	**32.84**	1.44	2.76	–	**39.97**	45.86	53.79	0.35
	SVR Chain	3.15	5.09	3.92	13.83	4.70	6.20	4.87	25.13	5.44	10.05	–	16.56	42.87	56.04	1.09
	SVR Multi	2.69	5.17	4.01	15.82	3.74	6.10	4.12	25.33	4.57	10.68	–	16.68	45.16	53.74	1.10
BE	CNN	3.82	5.55	4.84	6.48	5.90	23.46	10.93	7.83	20.35	8.31	–	–	49.71	47.76	2.53
	DNN	7.77	6.88	6.55	7.82	7.07	11.37	10.96	8.82	20.00	10.78	–	–	38.44	59.57	1.99
	RF	2.14	2.75	3.35	2.66	11.70	4.94	**33.21**	2.01	**30.47**	5.44	–	–	47.11	51.56	1.33
	SVR Chain	7.79	6.20	4.12	7.05	5.47	10.04	16.22	6.79	22.80	10.07	–	–	38.31	58.23	3.46
	SVR Multi	7.66	5.76	4.08	7.00	5.62	10.43	15.72	6.56	23.55	10.33	–	–	39.59	57.11	3.30

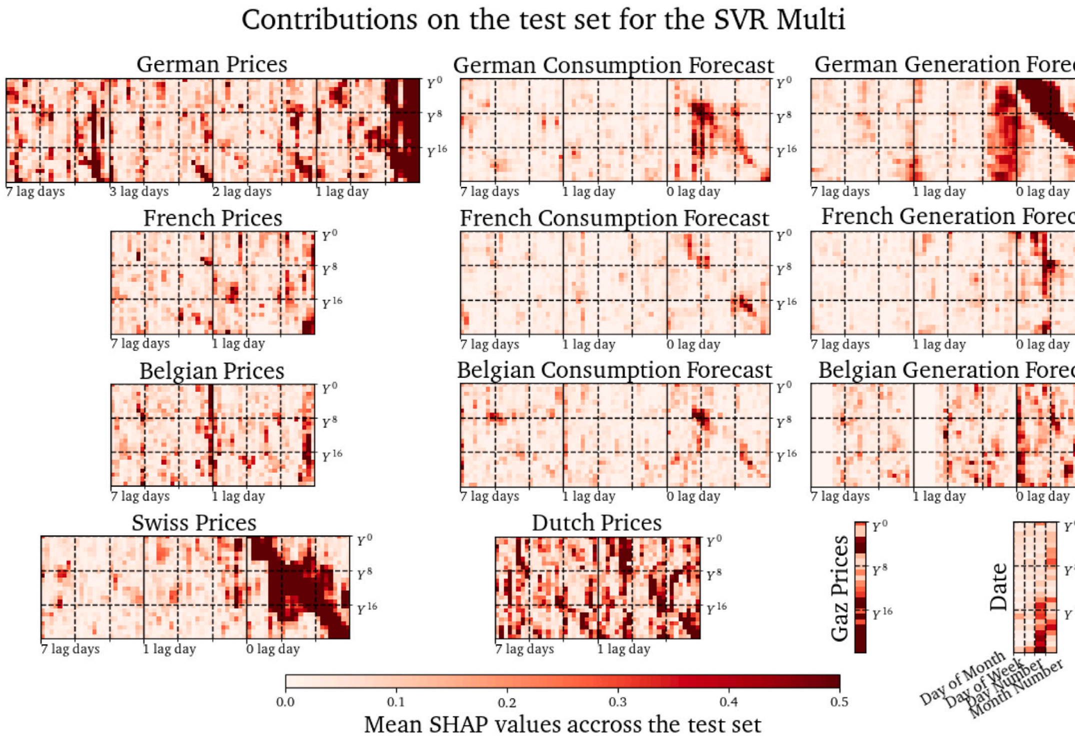

Fig. 4. Average feature contribution of the MultiSVR model for the German dataset. We observe diagonal lines of high contributions. This means that features of hour h contribute mainly for predicting the output o if $h = o$.

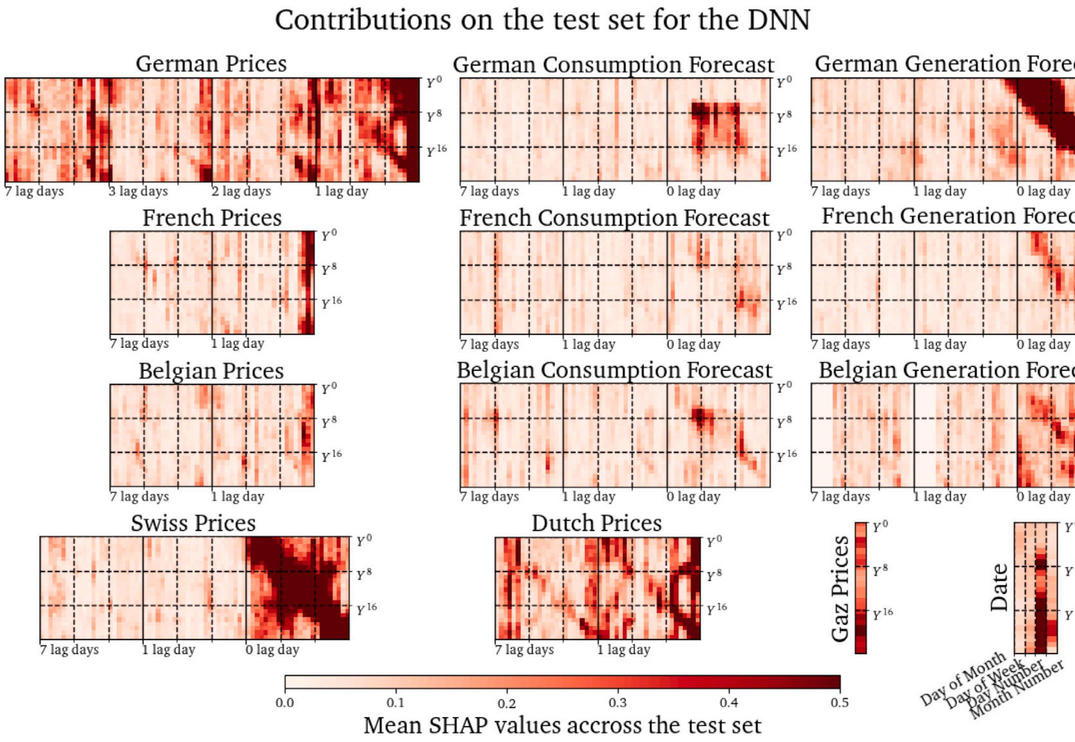

Fig. 5. Average feature contribution of the DNN model for the German dataset. We observe centered squares high contributions between hours from 8 am to 8 pm for French consumption and generation forecasts. The Neural Network models display peak-load related patterns in their feature contributions.

features, and attest of its capacity to integrate complex phenomenons. It also helps explaining the performance gap between all models. Similar figures for the CNN, ChainSVR and for other countries can be found in our repository.

We observe from Table 7 that both SVR and DNN models use foreign features for more than half of their total contribution (right-hand side columns). German renewable forecasts account for almost one fifth of the total contribution for predicting the German prices. This is the

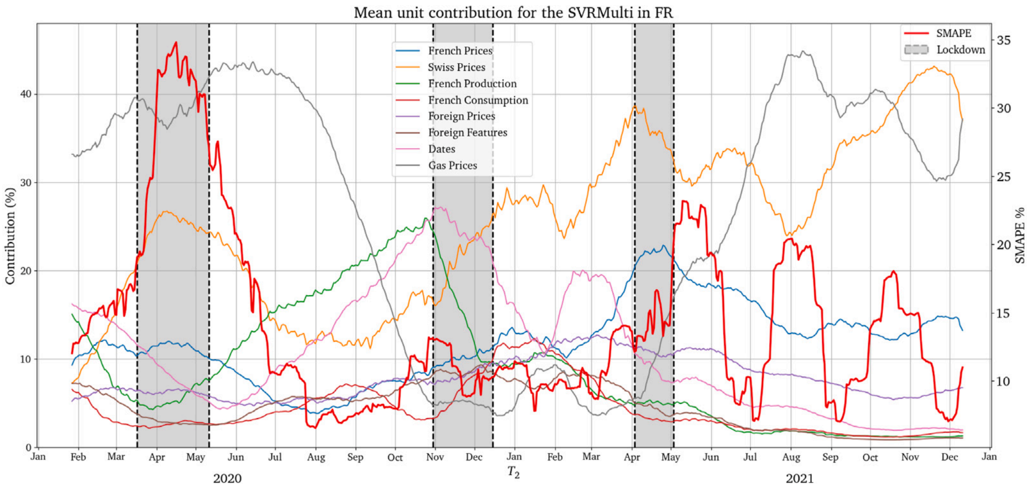

Fig. 6. Evolution of the daily mean unit contribution of each feature for the prediction of French prices over the period T_2 in the DNN model. The daily RMAE is displayed in red.

Table 8
Summary of average contributions per lag across all datasets for the first period T_2. Contributions are summed for all targets, all times and features for each lag.

	Model	D	D-1	D-2	D-3	D-7
FR	CNN	20.27	37.59	4.05	4.21	33.88
	DNN	46.71	29.83	2.41	2.32	18.73
	RF	73.81	21.05	0.30	0.34	4.49
	ChainSVR	42.34	31.42	2.72	2.63	20.89
	MultiSVR	44.87	30.74	2.85	2.12	19.42
DE	CNN	38.65	33.48	3.89	3.06	20.93
	DNN	43.53	30.01	3.50	2.92	20.04
	RF	54.12	40.13	0.49	0.39	4.87
	ChainSVR	36.28	37.52	2.92	2.92	20.36
	MultiSVR	38.62	37.12	2.55	2.48	19.24
BE	CNN	19.89	40.91	4.21	5.18	29.80
	DNN	28.39	39.72	3.99	3.89	24.01
	RF	21.55	64.20	1.25	1.31	11.69
	ChainSVR	29.55	43.65	3.90	3.27	19.63
	MultiSVR	29.61	43.72	3.93	3.30	19.44

highest contribution after Swiss and German prices and it is almost twice the contribution of the French generation forecast for predicting French prices (or Belgian generation forecast for predicting the Belgian prices) that reach 10% at maximum. We explain this observation by the difference in energy mix between these countries. Nuclear electricity is produced according to the pricing algorithm, while renewable energies are generated independently and determine market prices. As a result, models have to put more weight to those features. We also observe that the French consumption forecast contributes to predict French prices more than the German consumption forecast for German prices (and Belgian consumption forecast for the Belgian prices). It reflects the thermosensibility of the French consumption, making it more volatile and more determinant for setting the prices. Our studied models consider that this feature is more decisive in setting prices and gives it more weight. In addition, it is clear form Table 7 that Swiss prices accounts for an important part of the feature contribution for all datasets that contain them. They are not part of the EUPHEMIA algorithm and can therefore be used by market participants to create their order books. Owners of power plants can use these prices to plan their production. Thanks to cross-border energy flows, market players can also exchange energy from and to this country using its price as a reference. Swiss prices thus constitute a good overview of the price level that will be reached by its neighboring countries. Finally, from Table 8, we can conclude that features with two or three day lag contribute very little to the decision. They never contribute more than 5% of the total contribution. However, features with one or seven days lag are an important part of the model decision process. Different seasonalities can be observed in electricity prices. Among them, the weekly seasonality is one of the most important: due to weekends, the prices of the week before are generally more similar than the prices of the previous 2 or 3 days. In addition, prices are fixed daily by the EUPHEMIA algorithm, that uses as input the order books specific to the corresponding day. Thus, the data of the previous days, although similar to the current data, are not decisive in settling the prices. Contribution weights assigned to features with two and three day lags confirm that our models are also mostly based on current data.

Lastly, we focus on the daily average unit contribution of the features and their evolution along time period T_2. Fig. 6 displays these values for the best performing model for the French country: the MultiSVR. Gray rectangles delimit the periods of confinement, and we have displayed the SMAPE model in red. First, we observe the evolution of SMAPE during the three lockdowns. The model hardly adapts to the first lockdown (March 17, 2020–May 11, 2020) and the highest error is reached in the middle of it. The second lockdown (November 2020) shows no significant increase in errors overall. The third confinement (April 2021) corresponds to a decrease in SMAPE. We explain this evolution by two factors: (1) The first confinement was more brutal for the French market. Because industry has come to a standstill, prices and consumption have fallen. This is not the case for the other two lockdowns. (2) The first lockdown was a completely new situation for the model, which was not the case for the following. This also explains why the SMAPE drops after the first lockdown: the model has integrated several data samples from the confinement period and is able to adapt. Thus, the next two confinements are more easily dealt with. Next, we focus on the period following the first lockdown. On the market, this period was characterized by a slow French industry recovery and warm temperatures with consumption still below the standards. We see the Swiss and gas price contributions falling at the benefit of French production forecasts and date indicators. The model correctly balances the trade-off between French production forecasts and gas prices. Indeed, during this period, the French nuclear production fleet was enough to cover the consumption and no gas-fired plant was required. As a result, the SMAPE on this period is low. The third lockdown (May–December 2021) is opposite to the previous one. High volatility in gas prices, due in part to the economic recovery in China, is driving volatility in electricity prices, as shown in Fig. 1. In this context, let us look at the relationship between the contributions of Swiss and gas prices. These two features are most of the time the two most important characteristics and their movements are generally opposite: when one decreases, the other increases. Periods when the gas price contribution increases and the Swiss price contribution decreases are marked by a SMAPE spike. The model captures the importance of the gas price as shown by the strong increase in the contribution, but this price is so volatile that using gas price from two days ago results in a high error. A good way to avoid error spikes over this period would be to find a more reliable value than the EGSI index to estimate the price of gas two days before. Using the last value traded on the gas market could be an alternative. Lastly, we notice little variation in contribution among the other features. The French consumption forecast contributed more from January to March 2020 and from November 2020 to April 2021. During these winter periods, the model gave more weight to the consumption forecasts to reduce price temperature sensitivity. This is not the case during the winter of 2021. Very low French nuclear production, due to the maintenance of the power stations, obliges the gas park to ensure the balance between consumption and production. The proportion of gas-generated electricity in the total mix was so high that small variations in consumption did not affect prices. Indeed, all gas-fired power plants have the same marginal cost: the price of gas.

6. Synthesis, discussion and future work

In this section, we summarize our conclusions and observations from the results of our experiments. First, we see that including new features in the predictive dataset dramatically increases model performance. Among these added features, the most discriminating are the features without lag days: production and consumption forecasts, and Swiss prices. We believe that the Belgian dataset is more difficult to grasp as it lacks the forecasts for the period T_1 and the Swiss prices for the period T_2. We also observe that the feature contributions depend on the considered dataset. These differences reflect the specificities of the European market such as the temperature sensitivity of consumption in France, or the intermittency of production in Germany.

Second, we report significant inequalities in the performance of ML models. RF and CNN are not suitable for the EPF paradigm we are studying. These models incorrectly incorporate input features and therefore we cannot identify significant patterns in their contribution analysis. In contrast, DNN and SVR extract meaningful information from features and display diagonal and peak load patterns in their contributions. As a result, these models are better over the three considered countries and the two time periods. Further analysis of the contribution revealed that they are able to react to significant market changes by updating the weight of discriminating features such as gas price when necessary. Although this adaptation is not instantaneous and a short period of performance deterioration is observed, the models produce accurate predictions in new situations. For example, performance has increased during the second semester of 2021 for the French market even though prices are more volatile.

Due to the high computation times and the difficulty of acquiring new data, several experiments were left for future work. The integration of new EPF features such as coal, oil or carbon prices, or the use of more data from foreign countries such as Spain, Italy, Austria or Denmark could be considered as future work. Given the importance of Swiss and gas prices in the total contribution, it will also be interesting to include other prices without lag days available before the close of the EPEX market, such as EXAA prices or UK prices. Moreover, the available transfer capacities are essential to understand the cross-border energy flows that are necessary to explain the price differences between countries. Their inclusion in our datasets should increase the accuracy of the multi-country forecasting framework.

Finally, many other ML models could be tested, such as Gaussian processes, nearest neighbors or multi-kernel SVR. Regarding the significant contribution made by data with one or seven days lag, we believe that time series ML models such as recurrent neural networks, convolutional kernel random transformation models [86] or Dynamic Time warp models would challenge the benchmark state of the art. Additionally, we observed a slight degradation in performance while forecasting multiple countries at once. Managing European network topology using Graph Neural Network looks promising and will be our next challenge.

CRediT authorship contribution statement

Léonard Tschora: Conceptualization, Methodology, Implementation, Validation, Investigation of the results, Writing – original draft. **Erwan Pierre:** Investigation, Resources, Data curation, Writing, Funding acquisition. **Marc Plantevit:** Conceptualization, Investigation, Supervision, Writing, Funding acquisition. **Céline Robardet:** Conceptualization, Investigation, Supervision, Writing, Funding acquisition.

Declaration of competing interest

Leonard Tschora reports financial support was provided by National Association for Research and Technology.

Acknowledgment

This research has received funding from the ANRT (French National Association for Research and Technology).

References

[1] Amabile L, Bresch-Pietri D, El Hajje G, Labbé S, Petit N. Optimizing the self-consumption of residential photovoltaic energy and quantification of the impact of production forecast uncertainties. Adv Appl Energy 2021;2:100020. http://dx.doi.org/10.1016/j.adapen.2021.100020.

[2] Mei J, Zuo Y, Lee CH, Wang X, Kirtley JL. Stochastic optimization of multi-energy system operation considering hydrogen-based vehicle applications. Adv Appl Energy 2021;2:100031. http://dx.doi.org/10.1016/j.adapen.2021.100031.

[3] RTE. Bilan électrique 2019. Tech. rep., RTE; 2020.

[4] PCR. EUPHEMIA public description. Tech. rep., Price Coupling of Region; 2016.

[5] Narajewski M, Ziel F. Changes in electricity demand pattern in europe due to COVID-19 shutdowns. 2020.

[6] Suvarna M, Katragadda A, Sun Z, Choh YB, Chen Q, PS P, et al. A machine learning framework to quantify and assess the impact of COVID-19 on the power sector: An Indian context. Adv Appl Energy 2022;5:100078. http://dx.doi.org/10.1016/j.adapen.2021.100078.

[7] Krizhevsky A, Sutskever L, Hinto GE. ImageNet classification with deep convolutional neural networks. Tech. rep., University of Toronto; 2012.

[8] Zheng Y, Liu Q, Chen E, Ge Y, Zhao JL. Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. Tech. rep., University of Science and Technology of China Hefei and University of North Carolina Charlotte and City University of Hong Kong; 2015.

[9] Li Y, Yu R, Shahabi C, Liu Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. 2018.

[10] Vlahogianni EI, Karlaftis MG, Golias JC. Short-term traffic forecasting: Where we are and where we're going. Transp Res C 2014;43:3–19. http://dx.doi.org/10.1016/j.trc.2014.01.005.

[11] Lago J, Marcjasz G, De Schutter B, Weron R. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. Appl Energy 2021;293:116983. http://dx.doi.org/10.1016/j.apenergy.2021.116983.

[12] Weron R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. Int J Forecast 2014;30. http://dx.doi.org/10.1016/j.ijforecast.2014.08.008.

[13] Bagnall A, Bostrom A, Large J, Lines J. The great time series classification bake off: An experimental evaluation of recently proposed algorithms. Extended version. 2016.

[14] Ruiz AP, Flynn M, Large J, Middlehurst M, Bagnall A. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Discov 2021;35:401?449. http://dx.doi.org/10.1007/s10618-020-00727-3.

[15] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019;51(5):93:1–42. http://dx.doi.org/10.1145/3236009.

[16] Molnar C. Interpretable machine learning. Lulu. com; 2020.

[17] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artificial Intelligence Res 2021;70:245–317. http://dx.doi.org/10.1613/jair.1.12228.

[18] Covert I, Lundberg SM, Lee S. Explaining by removing: A unified framework for model explanation. J Mach Learn Res 2021;22:209:1–90.

[19] Islam MR, Ahmed MU, Barua S, Begum S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl Sci 2022;12(3):1353.

[20] Lundberg S, Lee S. A unified approach to interpreting model predictions. 2017, CoRR abs/1705.07874.

[21] Sansom D, Downs T, Saha T. Evaluation of support vector machine based forecasting tool in electricity price forecasting for Australian national electricity market participants. J Electr Electron Eng Australia 2003;22.

[22] Mei J, He D, Harley R, Habetler T, Qu G. A random forest method for real-time price forecasting in new york electricity market. In: 2014 IEEE PES general meeting| conference & exposition. IEEE; 2014, p. 1–5.

[23] Conejo A, Plazas M, Espinola R, Molina A. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. IEEE Trans Power Syst 2005;20:1035–42. http://dx.doi.org/10.1109/TPWRS.2005.846054.

[24] Pino R, Parreno J, Gomez A, Priore P. Forecasting next-day price of electricity in the spanish energy market using artificial neural networks. Eng Appl Artif Intell 2008;21(1):53–62. http://dx.doi.org/10.1016/j.engappai.2007.02.001.

[25] Miralles AR, Dorronsoro JR, Diaz J. Day-ahead price forecasting for the spanish electricity market. Tech. rep., Instituto de Ingenieria del Conocimiento and Universidad Autonoma de Madrid; 2019.

[26] Shi W, Wang Y, Chen Y, Ma J. An effective two-stage electricity price forecasting scheme. Electr Power Syst Res 2021;199:107416. http://dx.doi.org/10.1016/j.epsr.2021.107416.

[27] Shiri A, Afshar M, Rahimi-Kian A, Maham B. Electricity price forecasting using support vector machines by considering oil and natural gas price impacts. In: Mybook. 2015, p. 1–5. http://dx.doi.org/10.1109/SEGE.2015.7324591.

[28] Heijden TVD, Lago J, Palensky P, Abraham E. Electricity price forecasting in European day ahead markets: A greedy consideration of market integration. IEEE Access 2021;9:119954–66. http://dx.doi.org/10.1109/ACCESS.2021.3108629.

[29] Lago J, Ridder FD, Vrancx P, Schutter BD. Forecasting day-ahead electricity prices in Europe the importance of considering market integration. Tech. rep., Delft University of Technology and VITO-Energyville and Vrije Universiteit Brussel; 2017.

[30] Lago J, Ridder FD, Schutter BD. Forecasting day-ahead electricity prices Deep Learning approaches and empirical comparison of traditional algorithms. Tech. rep., Delft University of Technology and VITO-Energyville; 2018.

[31] Wang D, Luo H, Grunder O, Lin Y, Guo H. Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm. Appl Energy 2017;190:390–407. http://dx.doi.org/10.1016/j.apenergy.2016.12.134.

[32] Gunduz S, Ugurlu U, Oksuz I. Transfer learning for electricity price forecasting. 2020.

[33] Ziel F, Steinert R, Husmann S. Forecasting day ahead electricity spot prices: The impact of the exaa to other European electricity markets. Energy Econ 2015;51:430–44. http://dx.doi.org/10.1016/j.eneco.2015.08.005.

[34] Tan Z, Zhang J, Wang J, Xu J. Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. Appl Energy 2010;87(11):3606–10. http://dx.doi.org/10.1016/j.apenergy.2010.05.012.

[35] Che J, Wang J. Short-term electricity prices forecasting based on support vector regression and auto-regressive integrated moving average modeling. Energy Convers Manage 2010;51:1911–7. http://dx.doi.org/10.1016/j.enconman.2010.02.023.

[36] Contreras J, Espinola R, Nogales F, Conejo A. ARIMA models to predict next-day electricity prices. Power Eng Rev IEEE 2002;22:57. http://dx.doi.org/10.1109/MPER.2002.4312577.

[37] Crespo Cuaresma J, Hlouskova J, Kossmeier S, Obersteiner M. Forecasting electricity spot-prices using linear univariate time-series models. Appl Energy 2004;77:87–106. http://dx.doi.org/10.1016/S0306-2619(03)00096-5.

[38] Weron R, Misiorek A. Forecasting spot electricity prices with time series models. In: Proceedings of the European electricity market EEM-05 conference. 2005.

[39] Zhao J, Dong Z, Xu Z, Wong K. A statistical approach for interval forecasting of the electricity price. IEEE Trans Power Syst 2008;23:267–76. http://dx.doi.org/10.1109/TPWRS.2008.919309.

[40] Panapakidis IP, Dagoumas AS. Day-ahead electricity price forecasting via the application of artificial neural network based models. Appl Energy 2016;172:132–51. http://dx.doi.org/10.1016/j.apenergy.2016.03.089.

[41] Mosbah H, El-Hawary M. Hourly electricity price forecasting for the next month using multilayer neural network. Can J Electr Comput Eng 2015;39. http://dx.doi.org/10.1109/CJECE.2016.2586939.

[42] Keles D, Scelle J, Paraschiv F, Fichtner W. Extended forecast methods for day-ahead electricity prices applying artificial neural networks. Appl Energy 2016;SCI:218. http://dx.doi.org/10.1016/j.apenergy.2015.09.087.

[43] Anbazhagan S, Kumarappan N. Day-ahead deregulated electricity market price forecasting using recurrent neural network. IEEE Syst J 2013;7(4):866–72. http://dx.doi.org/10.1109/JSYST.2012.2225733.

[44] Khan Z, Fareed S, Anwar M, Naeem A, Gul H, Arif A, et al. Short term electricity price forecasting through convolutional neural network (CNN). In: BOOK. 2020.

[45] Ziel F, Steinert R. Electricity price forecasting using sale and purchase curves the X-Model. Energy Econ 2016;59:435–54. http://dx.doi.org/10.1016/j.eneco.2016.08.008.

[46] Schnürch S, Wagner A. Machine learning on EPEX order books insights and forecasts. 2019.

[47] Kulakov S. X-model: further development and possible modifications. 2019.

[48] Barta G, Nagy GBG, Kazi S, Henk T. GEFCOM 2014–p-robabilistic electricity price forecasting. In: Neves-Silva R, Jain LC, Howlett RJ, editors. Intelligent decision technologies. Cham: Springer International Publishing; 2015, p. 67–76.

[49] Juban R, Ohlsson H, Maasoumy M, Poirier L, Kolter JZ. A multiple quantile regression approach to the wind, solar, and price tracks of gefcom2014. Int J Forecast 2016;32(3):1094–102. http://dx.doi.org/10.1016/j.ijforecast.2015.12.002.

[50] Gaillard P, Goude Y, Nedellec R. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int J Forecast 2016;32(3):1038–50. http://dx.doi.org/10.1016/j.ijforecast.2015.12.001.

[51] Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. Int J Forecast 2016;32(3):896–913. http://dx.doi.org/10.1016/j.ijforecast.2016.02.001.

[52] Demir S, Mincev K, Kok K, Paterakis NG. Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. Appl Energy 2021;304:117695. http://dx.doi.org/10.1016/j.apenergy.2021.117695.

[53] Nowotarski J, Raviv E, Trück S, Weron R. An empirical comparison of alternative schemes for combining electricity spot price forecasts. Energy Econ 2014;46:395–412. http://dx.doi.org/10.1016/j.eneco.2014.07.014.

[54] Nan F, Bordignon S, Bunn D, Lisi F. The forecasting accuracy of electricity price formation models. Int J Energy Stat 2014;2:1–26. http://dx.doi.org/10.1142/S233568041450001X.

[55] Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 2021;296:103473.

[56] Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: NEURIPS. 2017, p. 4765–74.

[57] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R, editors. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016, p. 1135–44. http://dx.doi.org/10.1145/2939672.2939778.

[58] Shapley LS. 17. A value for n-person games. Princeton University Press; 1953.

[59] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. Google; 2016.

[60] He K, Zhang X, Ren S, Sun J. Deep residual networks for image recogition. Tech. rep, Microsoft; 2015, https://arxiv.org/abs/1512.03385.

[61] Batal I, Sacchi L, Bellazzi R, Hauskrecht M. Multivariate time series classification with temporal abstractions. Int J Artif Intell Tools Archit Languages Algorithms 2009;22:344–9.

[62] Orsenigo C, Vercellis C. Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. Pattern Recognit 2010;43:3787–94. http://dx.doi.org/10.1016/j.patcog.2010.06.005.

[63] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. Data Min Knowl Discov 2019;33(4):917–63. http://dx.doi.org/10.1007/s10618-019-00619-1.

[64] Kampouraki A, Manis G, Nikou C. Heartbeat time series classification with support vector machines. IEEE Trans Inf Technol Biomed 2009;13(4):512–8. http://dx.doi.org/10.1109/TITB.2008.2003323.

[65] Rakotomamonjy A, Bach F, Canu S, Grandvalet Y. SimpleMKL. J Mach Learn Res 2008;9:2491–521.

[66] Che J, Wang J. Short-term load forecasting using a kernel-based support vector regression combination model. Appl Energy 2014;132:602–9. http://dx.doi.org/10.1016/j.apenergy.2014.07.064.

[67] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.

[68] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.

[69] Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series. European J Oper Res 2005;160(2):501–14.

[70] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. Heliyon 2018;4(11):e00938. http://dx.doi.org/10.1016/j.heliyon.2018.e00938.

[71] Chen Y, Zhang D. Theory-guided deep-learning for electrical load forecasting (TgDLF) via ensemble long short-term memory. Adv Appl Energy 2021;1:100004. http://dx.doi.org/10.1016/j.adapen.2020.100004.

[72] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. In: The handbook of brain theory and neural networks. Cambridge, MA, USA: MIT Press; 1998, p. 255–8.

[73] Badrinarayanan V, Handa A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. 2015.

[74] Zheng Y, Liu Q, Chen E, Ge Y, Zhao J. Time series classification using multi-channels deep convolutional neural networks. In: WAIM 2014. LNCS, vol. 8485. 2014, p. 298–310.

[75] Borovykh A, Bohte S, Oosterlee CW. Conditional time series forecasting with convolutional neural networks. Tech. rep., Universita di Bologna and Centrum Wiskunde and Delft University of Technology; 2018.

[76] Cheng H-Y, Kuo P-H, Shen Y, Huang C-J. Deep convolutional neural network model for short-term electricity price forecasting. 2020.

[77] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015, Software available from tensorflow.org.

[78] Diebold F, Mariano R. Comparing predictive accuracy. J Bus Econom Statist 1992;20:134–44. http://dx.doi.org/10.1080/07350015.1995.10524599.

[79] Diebold F. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. J Bus Econom Statist 2012;33. http://dx.doi.org/10.2139/ssrn.2144210.

[80] Uniejewski B, Weron R, Ziel F. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans Power Syst 2018;33(2):2219–29. http://dx.doi.org/10.1109/TPWRS.2017.2734563.

[81] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(2).

[82] Sheppard D. Gas shortages: what is driving Europe's energy crisis?. 2021.

[83] Figueroa MG, Cartea A. Pricing in electricity markets: A mean reverting jump diffusion model with seasonality. Appl Math Finance 2015;12(4):313–35.

[84] Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. 2018, CoRR abs/1805.10820.

[85] Lundberg SM, Erion GG, Chen H, DeGrave AJ, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

[86] Dempster A, Petitjean F, Webb GI. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. 2019, CoRR abs/1910.13051.