

ORIGINAL RESEARCH

Open access

Explainability-Driven Trust Framework for Electricity Price Forecasting Models

Kevin O'Brien^{1*}, Liam Murphy¹, Sean Doyle²

Abstract

Sophisticated machine learning (ML) techniques have demonstrated superior performance over conventional methods across numerous forecasting tasks, particularly in electricity price forecasting (EPF). Nevertheless, the predictive capability of these ML models deteriorates markedly when encountering input data that differs significantly from the examples observed during training. This phenomenon frequently occurs in EPF due to shifts in market conditions, such as surges in fuel costs, higher integration of renewable energy sources, or modifications in regulatory frameworks and operational rules. While the decline in accuracy on unfamiliar data is problematic, an even greater difficulty lies in the inability to anticipate when the model will produce unreliable outputs. This lack of foresight exposes electricity market participants—including bidding agents and retailers—to considerable financial risks stemming from erroneous price predictions. Consequently, it is critical to determine whether a given model prediction can be relied upon in real time. To address this need, the present study introduces a trust evaluation algorithm for EPF applications grounded in explainable artificial intelligence (XAI) methods. The proposed algorithm computes trust scores that indicate the reliability of each new prediction generated by the model. These scores are constructed in two phases: first, a preliminary score is derived from the correlation between local and global explanations; second, the score is refined using SHapley Additive exPlanations (SHAP) values for individual features. Such score-based interpretations are more accessible and actionable for EPF practitioners, such as asset managers and traders, than conventional feature-based visualizations. The effectiveness of the algorithm is assessed using datasets from the Italian electricity market and the Electric Reliability Council of Texas (ERCOT). Empirical results reveal that the method achieves over 85% accuracy in recognizing reliable predictions when the incoming data distribution closely matches the training distribution. Under conditions of distributional shift, the algorithm maintains a comparable level of accuracy in correctly flagging unreliable predictions.

Keywords Explainability, Electricity price forecasting, EPF, Explainable AI model, XAI, SHAP

*Correspondence:

Kevin O'Brien

kevin.obrien@ucd.ie

¹ Department of Smart Grid Analytics, Faculty of Engineering, University College Dublin, Dublin, Ireland

² Department of Energy Systems Optimization, Faculty of Engineering, Trinity College Dublin, Dublin, Ireland

Introduction

Electricity prices have become increasingly volatile as the share of intermittent renewable energy sources continues to grow [1, 2]. Because bidding strategies of market participants rely heavily on accurate electricity price forecasts, heightened uncertainty in price movements exposes them to substantial financial risks [3, 4]. In recent years, advances in deep learning have led to the development of more effective machine learning (ML) and artificial intelligence (AI) algorithms for a wide range of power system applications [5–7], including electricity price forecasting (EPF) [8–13]. In specific contexts, these algorithms offer clear benefits over traditional techniques in terms of computational efficiency, robustness to noise, and overall forecasting accuracy [14].

Despite these advantages, a fundamental limitation persists: the internal complexity of modern ML models often renders their decision-making processes opaque. As a result, users in the energy sector may struggle to understand or trust the outputs and recommendations produced by such systems, which restricts their widespread adoption. This challenge is particularly acute in domains demanding high reliability, a hallmark of the energy industry [15], and is especially relevant for EPF tasks. In response, novel methodologies and principles have emerged in recent years to enhance the transparency of ML models, collectively known as Explainable Artificial Intelligence (XAI).

XAI techniques have found application in various power system contexts [16, 17]. Among the most widely adopted XAI tools in EPF is SHapley Additive exPlanations (SHAP), which draws on game-theoretic principles to elucidate model predictions by quantifying the contribution of each input feature to a specific forecast [18]. Several studies have leveraged SHAP to validate the importance of newly introduced features in EPF models [19, 20]. Similarly, research on the German electricity market used SHAP to demonstrate the pivotal role of generation in determining prices on the NordPool exchange [21]. Another contribution proposed a comprehensive framework for constructing EPF models while incorporating SHAP analysis to facilitate greater human-in-the-loop involvement [22]. Additionally, an LSTM-based EPF model incorporating a multi-head self-attention mechanism was developed to highlight the temporal relevance of different input variables [23]. Collectively, these works emphasize feature-importance explanations for EPF models.

Explanations for EPF-related ML models can be tailored to two primary user groups: AI experts and EPF practitioners. AI experts, typically model developers and researchers, benefit from feature-importance analyses that aid in model debugging and architectural refinement. In contrast, EPF practitioners—such as asset managers and bidding agents—require concise, post-hoc explanations [24] that directly support operational decision-making, including bidding strategies and scheduling. Feature-based visualizations offer limited practical value for these users. Instead, score-based explanations, particularly trust scores, better meet their needs by enabling rapid assessment of whether an individual model prediction should be trusted. When trust scores fall below a predefined threshold, they signal a potential deterioration in model accuracy, prompting users to adjust their strategies accordingly. The practical utility of such score-based explanations for asset managers is illustrated conceptually in **Figure 1**, which also highlights the core contribution of the current work.

While existing literature has predominantly concentrated on generating explanations for AI experts, this paper introduces a trust evaluation algorithm rooted in XAI techniques specifically designed for EPF users. The algorithm produces a trust score for every prediction issued by the trained ML model. Score generation occurs in two sequential stages: an initial coarse score is computed based on the correlation between local and global explanations, followed by a refinement step that incorporates the SHAP values of relevant features. A further contribution of this study is the provision of a precise definition for the prediction outcome vectors generated by the proposed algorithm, addressing a common ambiguity in XAI research [25]. Through empirical evaluation, the work demonstrates how these trust scores can effectively identify instances in which model accuracy declines over time. The proposed methodology is validated using real-world data from the Electric Reliability Council of Texas (ERCOT) and the Italian electricity market.

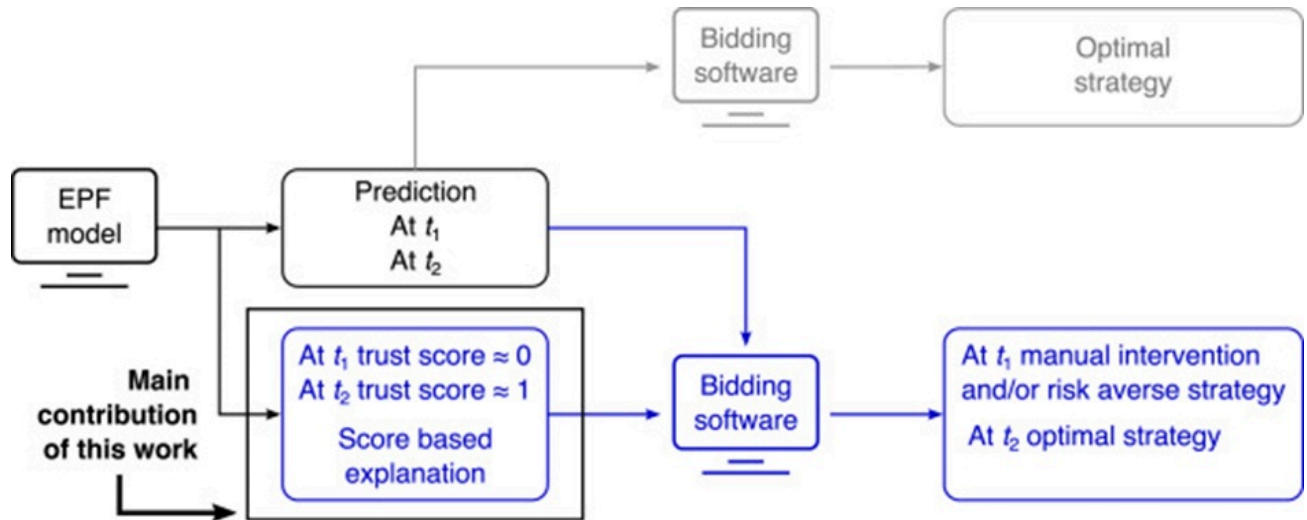


Figure 1. Schematic illustration of the score-based explanation approach.

Technical Background

EPF model

Let the Electricity Price Forecasting (EPF) task be modeled by a machine learning function $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$. This model is trained using the dataset $D = [X, F]$. At each time step t , the input vector, $X_t^{EPF} \in \mathbb{R}^m$ consists of the following components:

$$X_t^{EPF} = \left[t, P_{daily,t}^{avg}, P_{t-i}, \dots, P_{t-1}, P_{t-1}^{NZ_1}, \dots, P_{t-1}^{NZ_j}, P_{t-1}^{NC_1}, \dots, P_{t-1}^{NC_k} \right]. \quad (1)$$

The features include the hour of the day t , the average daily electricity price $P_{daily,t}^{avg}$, past electricity prices $[P_{t-i}, \dots, P_{t-1}]$, locational marginal prices from various zones within the country $[P_{t-1}^{NZ_1}, \dots, P_{t-1}^{NZ_j}]$, and electricity prices from neighboring countries $[P_{t-1}^{NC_1}, \dots, P_{t-1}^{NC_k}]$. Here, indices i , j , and k denote the number of lagged prices, the number of domestic price zones, and the number of neighboring countries considered, respectively.

The corresponding output vector $Y_t^{EPF} \in \mathbb{R}^n$ is defined as:

$$Y_t^{EPF} = [P_t, P_{t+1}, P_{t+2}, \dots, P_{t+n-1}], \quad (2)$$

Where P_t denotes the electricity price at time t , n is the prediction horizon length, and the remaining terms represent the subsequent $(n-1)$ -step-ahead forecasts.

The input vector described above corresponds to the Convolutional Neural Network (CNN) architecture used for the Italian Electricity Market. For the discussions in this section, this formulation serves as the reference; however, the underlying concepts apply equally to other input structures and model types. This generality is verified experimentally by applying it to the ERCOT market using a Long Short-Term Memory (LSTM) model. For the ERCOT case, the input vector is given by:

$$X_t^{EPF} = \left[t, P_{t-i}, \dots, P_{t-1}, P_{t-1}^{NZ_1}, \dots, P_{t-1}^{NZ_j}, P_{t-1}^{AM_1}, \dots, P_{t-1}^{AM_k} \right]. \quad (3)$$

Where $P_{t-1}^{AM_k}$ represents the ancillary service market prices at time t-1. In this setup, k = 3, accounting for up-regulation, non-spin, and responsive reserve service prices. Note that the ERCOT market input data is normalized, whereas the Italian market data is not. The proposed algorithm performs effectively in both scenarios.

Local fit and density principles

Previous work [26] identified two primary factors influencing the reliability of machine learning models. The first concerns whether the current input instance resembles the data observed during training. This relates to the density principle, which examines the distribution of training samples surrounding the input point. When the input lies in a low-density region or is an outlier, the model must extrapolate, often leading to reduced accuracy under out-of-distribution (OOD) conditions. The second factor assesses the model's accuracy, assuming the input is similar to training instances. This corresponds to the local fit principle, which measures how well the model fits the training samples nearest to the given input. This principle is relevant when the input belongs to the independent and identically distributed (IID) portion of the data space. Several reliability evaluation techniques grounded in these density and local fit concepts have been introduced in the literature [27-29].

PFI, SHAP, and LR scores of the EPF model

Permutation Feature Importance (PFI) scores $\phi_{XEPF,p}$ offer a global explanation of the EPF model and are expressed as:

$$\phi_{XEPF,p} = \epsilon_{XEPF} - \epsilon_{XEPF,p}, \quad (4)$$

Where ϵ_{XEPF} is the mean squared error on the training set, and $\epsilon_{XEPF,p}$ denotes the error after permuting the pth feature in the input vector.

SHAP values $\phi_{x_{tp}}$ provide a local interpretability measure for the EPF model, quantifying the contribution of the pth feature x_{tp} within the input vector X_t^{EPF} at instance t. It is formulated as:

$$\phi_{x_t^p} = \frac{1}{Z} \sum_{z=1}^Z \hat{f}_{X_{S \cup x_t^p}^z} - \hat{f}_{X_S^z}, \quad (5)$$

Where S denotes the subset of features drawn from the input vector X_t^{EPF} , Z represents the total count of coalition vectors evaluated, and z stands for each binary coalition vector that includes some features while excluding others. Assume the pth feature corresponds to P_{t-1NZj} , meaning $x_{tp} = P_{t-1NZj}$. The input vector corresponding to a random coalition vector z (with the pth feature x_{tp} included) is expressed as:

$$X_{S \cup P_{NZj}}^z = \left[t, P_{daily,t}^{avg}, P_{t-i}, \dots, P_{t-1}^a, P_{t-1}^{NZ_1,a}, \dots, P_{t-1}^{NZ_j}, P_{t-1}^{NC_1,a}, \dots, P_{t-1}^{NC_k,a} \right]. \quad (6)$$

In this formulation, features absent according to z are marked with a superscript a. These missing features are either replaced with randomly sampled values from the background dataset or with the feature's mean. The input vector in which the feature x_{tp} is excluded is formulated as:

$$X_S^z = \left[t, P_{daily,t}^{avg}, P_{t-i}, \dots, P_{t-1}^a, P_{NZ_1}^a, \dots, P_{NZ_j}^a, P_{NC_1}^a, \dots, P_{NC_k}^a \right]. \quad (7)$$

Linear regression (LR) models are naturally interpretable. In this study, LR scores are defined as:

$$\Phi_{x_t^p} = \beta_t^p x_t^p, \quad (8)$$

Where β_t^p indicates the linear regression coefficient associated with the p th feature x_t^p at time instance t .

Aggregate performance metrics

The mathematical formulation for the Mean Absolute Error concerning the d th component of the output Y_t^{EPF} is presented as:

$$MAE_d = \frac{1}{T} \sum_{t=1}^T |\hat{y}_d^{EPF} - y_d^{EPF}|, \quad (9)$$

Where $\hat{y}_{d,t}^{EPF}$ stands for the d th element of the model's predicted output at instance t , y_d^{EPF} signifies the true electricity price, and T denotes the total number of observations in the test dataset. The absolute error itself is written as:

$$\epsilon_d^{abs} = |\hat{y}_d^{EPF} - y_d^{EPF}|. \quad (10)$$

Likewise, the expression for Mean Absolute Percentage Error takes the form:

$$MAPE_d = \frac{1}{T} \sum_{t=1}^T \left| \frac{\hat{y}_d^{EPF} - y_d^{EPF}}{y_d^{EPF}} \right| \times 100, \quad (11)$$

While the absolute percentage error is given by:

$$\epsilon_d^{percent} = \left| \frac{\hat{y}_d^{EPF} - y_d^{EPF}}{y_d^{EPF}} \right| \times 100. \quad (12)$$

Definition of good and bad predictions

Unlike classification tasks, which involve discrete outcomes and thus allow straightforward correctness judgments, regression tasks produce continuous values, making correctness judgments more difficult. Assigning binary right/wrong labels to continuous forecasts is therefore not feasible. To address this, the present work introduces explicit criteria for classifying predictions as "good" or "bad."

Definition 1

Let the prediction outcome vector be defined as:

$$\gamma_t^{MAE} = [\gamma_t^{MAE_1}, \gamma_t^{MAE_2}, \dots, \gamma_t^{MAE_d}, \dots, \gamma_t^{MAE_n}] \quad (13)$$

Where:

$$\gamma_t^{MAE_d} = \begin{cases} 1, & \text{if } \epsilon_{d,t}^{abs} \leq MAE_d, \\ 0, & \text{otherwise,} \end{cases}$$

$$\forall d = 1, 2, \dots, n$$

In this context, $YMAEd_t = 1$ signifies a “good” prediction and $YMAEd_t = 0$ denotes a “bad” prediction according to the MAEd criterion.

Definition 2

Let the prediction outcome vector $YMAPE_t$ be defined as:

$$\gamma_t^{MAPE} = [\gamma_t^{MAPE_1}, \gamma_t^{MAPE_2}, \dots, \gamma_t^{MAPE_d}, \dots, \gamma_t^{MAPE_n}] \quad (14)$$

Where:

$$\gamma_t^{MAPE_d} = \begin{cases} 1, & \text{if } \epsilon_{d,t}^{abs} \leq MAPE_d, \\ 0, & \text{otherwise,} \end{cases}$$

$$\forall d = 1, 2, \dots, n$$

Here, $YMAPEd_t = 1$ corresponds to a “good” prediction and $YMAPEd_t = 0$ corresponds to a “bad” prediction based on the MAPEd criterion.

Materials and Methods

This section first lays out the mathematical basis of the proposed algorithm. It then provides an overview of the complete procedure.

Proposed trust algorithm—general

The Trust algorithm takes feature importance information generated by local explanations (via SHAP) and global explanations (via PFI and LR) and converts it into an easy-to-understand trust score. To appreciate how these explanation metrics and their interrelationships can signal the reliability of a forecast, it is useful to begin by considering the sources of prediction error. Such errors generally stem from two causes: (i) input features that differ substantially from patterns observed during model training (the density principle); and (ii) the model's inability to fully capture certain characteristics of the underlying stochastic process present in the training data (the local fit principle). Prediction error can be expressed as:

$$\epsilon_t^{abs} = |\hat{Y}_t^{EPF} - Y_t^{EPF}| = |\hat{f}(X_t^{EPF}) - Y_t^{EPF}| = \xi(Y^{EPF}, X_t^{EPF}, \hat{f}). \quad (15)$$

Eq. 15 shows that the absolute error vector ϵ_{abs} depends on the actual electricity price Y_t , the input features X_t^{EPF} , and the trained model $\hat{f}(X_t^{EPF})$. Although future electricity prices Y_t^{EPF} remain unknown, the input features and the model are

both available. SHAP values along with their correlations to PFI and LR explanations are shaped by X_t^{EPF} and $\hat{f}(X_t^{EPF})$. These elements can therefore serve as proxies for estimating prediction quality, i.e., whether significant error is likely present. This insight underpins the entire Trust algorithm.

The algorithm proceeds in two distinct stages. The first stage primarily addresses the local fit principle, while the density principle is incorporated in both stages.

First stage of the proposed algorithm

Instead of measuring distance, the first stage evaluates density through similarity. Correlation statistics are used to gauge how closely the current input resembles the training data. Given that the features are time series with pronounced seasonal behavior, correlations are calculated between the incoming input and the corresponding period in the historical training set. Each correlation test produces a coefficient and an associated p-value. The correlation is deemed statistically meaningful only when the p-value falls below 0.08 [30]. Averaging these correlation coefficients then provides a quantitative measure of similarity-based density. Strong, statistically significant correlations imply that the input shares characteristics with examples the model encountered during training.

Next, the local fit principle is examined by correlating the SHAP values for the current prediction period against three references:

- (i) SHAP values from the identical period in the previous year
- (ii) PFI scores
- (iii) LR scores

Reasoning for correlation-based analysis: Correlations reveal whether the model's current behavior aligns with the patterns it learned from training data. Consider two hypothetical models. Model $\hat{f}_1$ has successfully captured all important traits of the stochastic process. Model $\hat{f}_2$, trained on the same dataset, has missed one or more key traits.

From Eq. 4, PFI scores φ_{xEPF} can be viewed as a function Γ of the training set $[XEP, EPF]$ and the fitted model $\hat{f}$:

$$\varphi_{xEPF} = \Gamma(X^{EPF}, Y^{EPF}, \hat{f}). \quad (16)$$

This means PFI reflects knowledge derived from both actual process outcomes and the model's responses during training. Conversely, SHAP scores ϕ_{xp} , per Eq. 5, depend only on the current input vector X_t^{EPF} and the model $\hat{f}$:

$$\phi_{x_t^p} = \gamma(X_t^{EPF}, \hat{f}). \quad (17)$$

Consequently, SHAP explanations do not incorporate any process information beyond what is already embedded in the PFI scores.

In the case of a well-trained model ($\hat{f}_1$), PFI and SHAP should produce consistent feature rankings and relative magnitudes. This alignment typically yields a high Pearson correlation coefficient r_{SHAP} , close to 1, accompanied by a statistically significant p-value [31]. The coefficient is computed as:

$$r_{SHAP,PFI} = \frac{\sum (\varphi_{xEPF,p} - \bar{\varphi}_{xEPF,p}) (\phi_{x_t^p} - \bar{\phi}_{x_t^p})}{\sqrt{\sum \left(\varphi_{xEPF,p}^2 - \bar{\varphi}_{xEPF,p}^2 \right) \sum \left(\phi_{x_t^p}^2 - \bar{\phi}_{x_t^p}^2 \right)}}. \quad (18)$$

For the poorly trained model (72), PFI may still capture additional process characteristics by including output values. In contrast, SHAP remains limited to what can be inferred from the input and the model alone. As a result, the correlation in Eq. 18 tends to weaken — the coefficient may approach zero, the p-value may exceed the threshold, or both may occur. These patterns allow correlations to indicate whether the model is performing reliably for a given input, assuming the input lies within the training distribution (IID).

Implementation steps of the first stage: This stage evaluates correlations between the current local SHAP values and three benchmarks — previous-year SHAP values for the same period, global PFI scores, and LR scores. While PFI and LR offer broad, model-level insights, previous-year SHAP values provide localized historical context. The resulting correlation statistics help determine whether the model's response to the current input matches its expected behavior on in-distribution data. The correlation between current and previous-year SHAP scores is given by:

$$r_{SHAP,SHAP} = \frac{\sum \left(\phi_{x_t^p}^{old} - \bar{\phi}_{x_t^p}^{old} \right) (\phi_{x_t^p} - \bar{\phi}_{x_t^p})}{\sqrt{\sum \left(\left(\phi_{x_t^p}^{old} \right)^2 - \left(\bar{\phi}_{x_t^p}^{old} \right)^2 \right) \sum \left(\phi_{x_t^p}^2 - \bar{\phi}_{x_t^p}^2 \right)}}, \quad (19)$$

Where ϕ_{oldp} denotes the SHAP values from the corresponding period one year earlier. The correlation between SHAP and LR scores is expressed as:

$$r_{SHAP,LR} = \frac{\sum (\phi_{x_t^p} - \bar{\phi}_{x_t^p}) (\phi_{x_t^p} - \bar{\phi}_{x_t^p})}{\sqrt{\sum \left(\phi_{x_t^p}^2 - \bar{\phi}_{x_t^p}^2 \right) \sum \left(\phi_{x_t^p}^2 - \bar{\phi}_{x_t^p}^2 \right)}}. \quad (20)$$

The correlation between the current input vector X_t^{EPF} and the matching training data from the same seasonal period X_t^{EP} , is:

$$r_{X_t^{EPF}, X_t^{EPF,O}} = \frac{\sum \left(X_t^{EPF} - \bar{X}_t^{EPF} \right) \left(X_t^{EPF,O} - \bar{X}_t^{EPF,O} \right)}{\sqrt{\sum \left(\left(X_t^{EPF} \right)^2 - \left(\bar{X}_t^{EPF} \right)^2 \right) \sum \left(\left(X_t^{EPF,O} \right)^2 - \left(\bar{X}_t^{EPF,O} \right)^2 \right)}}. \quad (21)$$

Because these calculations use data from analogous seasonal periods, they produce correlation vectors that are subsequently averaged. The associated p-values are labeled $p_{SHAP,P}$, $p_{SHAP,PFI}$, $p_{SHAP,LR}$, and $p_{XEPF,XEPF,O}$ respectively.

Coarse form of Trust Score: The first stage computes a preliminary (coarse) trust score T_c . When $r_{XEPF_t, XEPF_{Ot}} \geq 0.7$, the score is determined as follows:

$$T_c = \begin{cases} r_{SHAP, SHAP}, & \text{if } p_{SHAP, SHAP} \leq 0.08, \\ r_{SHAP, PFI}, & \text{if } p_{SHAP, SHAP} > 0.08. \end{cases} \quad (22)$$

If both $p_{SHAP, SHAP} \leq 0.08$ and $p_{SHAP, PFI} \leq 0.08$ hold, then T_c is set to the larger of $r_{SHAP, SHAP}$, and $r_{SHAP, PFI}$. If both of these correlations lack statistical significance while $p_{SHAP, LR} \leq 0.08$, then:

$$T_c = \begin{cases} r_{SHAP, LR}, & \text{if } p_{SHAP, LR} \leq 0.08, \\ r_{X_t^{EPF}, X_t^{EPF, O}}, & \text{if } p_{SHAP, LR} > 0.08. \end{cases} \quad (23)$$

However, when $r_{XEPF_t, XEPF_{Ot}} \leq 0.7$ or the input-training correlation is not statistically significant, the rules above still apply except that $r_{SHAP, SHAP}$ is excluded. Low similarity between the current input and historical training data makes local SHAP-to-SHAP correlations less informative. All correlations and p-values in this stage are calculated using Pearson's method, selected based on prior observations of the dataset's characteristics. Alternative correlation measures (Spearman or Kendall) may be substituted depending on the nature of the data.

Second stage of the proposed algorithm

The second stage refines the coarse score T_c into the final trust score T by further applying the density principle through SHAP values. Specifically, it considers two properties of the SHAP values: (i) the highest individual SHAP value $\phi_{\max}$ across all features, and (ii) the total sum of SHAP values ϕ_{sum} . These are compared against a general threshold ϕ_{lim} to adjust T_c accordingly.

Reasoning for incorporating SHAP values: SHAP scores can help diagnose forecast reliability by highlighting how similar or dissimilar the current inputs are to the training distribution. Two contrasting situations are relevant: inputs that closely resemble training examples versus inputs that deviate markedly in magnitude, relationships, or both.

The key difference lies in the magnitude of $\hat{X}_{ZSuxpt} - \hat{X}_{ZS}$ from [Eq. 5](#). This difference tends to be modest when inputs are similar to training data and large when they are dissimilar. Returning to the marginal contribution example for feature PNZ_j with background samples q_1 to q_s , suppose in a similar case the observed value is $P_{1NZ_j} = q$ (within the background range). In the dissimilar case, it becomes $P_{2NZ_j} = q + \Delta q$, where Δq represents a substantial deviation.

For the similar-input case, the relevant vectors appear as in [Eqs. 24 and 25](#).

$$X_{S \cup P_{NZ_j}}^{z^1} = [t, P_{daily, t}^{avg}, P_{t-i}, \dots, P_{t-1}^a, P_{NZ_1, t-1}^a, \dots, q_1, P_{NC_1, t-1}^a, \dots, P_{NC_k, t-1}^a], \quad (24)$$

$$X_S^{z^1} = [t, P_{daily, t}^{avg}, P_{t-i}, \dots, P_{t-1}^a, P_{NZ_1, t-1}^a, \dots, q_1^a, P_{NC_1, t-1}^a, \dots, P_{NC_k, t-1}^a]. \quad (25)$$

Here, q_1 could be either any one of the values from $[q_1, q_2, \dots, q_s, q]$ or q_{1avg} , the mean of all values, as shown by

$$q_{avg}^1 = \frac{q_1 + q_2 + \dots + q_s + q}{s+1}. \quad (26)$$

In the first scenario, when the input vector closely resembles the training data, the feature value q remains near q_{a1} . Consequently, as indicated by [Eqs. 24-26](#), the vectors X_{z1S} and $X_{z1S \cup PNZj}$ become quite similar. This results in a small difference between $\hat{\phi}(X_{z1S \cup PNZj})$ and $\hat{\phi}(X_{z1S})$. As a result, the magnitude of the corresponding SHAP score, expressed as:

$$\phi_{x_i^1}^1 = \frac{1}{Z} \left(\sum_{z=1}^Z \hat{f}(X_{S \cup PNZj}^{z1}) - \hat{f}(X_S^{z1}) \right), \quad (27)$$

It will be low.

In contrast, for the second scenario, $X_{S2 \cup PNZj}$ is defined by [Eq. 28](#):

$$X_{S \cup PNZj}^{z1} = [t, P_{daily,t}^{avg}, P_{t-i}, \dots, P_{t-1}^a, P_{NZ_1,t-1}^a, \dots, q + \Delta q, P_{NC_1,t-1}^a, \dots, P_{NC_k,t-1}^a], \quad (28)$$

Whereas X_{S2} is given by [Eq. 29](#).

$$X_S^{z1} = [t, P_{daily,t}^{avg}, P_{t-i}, \dots, P_{t-1}^a, P_{NZ_1,t-1}^a, \dots, q_2^a, P_{NC_1,t-1}^a, \dots, P_{NC_k,t-1}^a]. \quad (29)$$

Here, q_{a2} may correspond to any individual value from the set $[q_1, q_2, \dots, q_s, q + \Delta q]$, or it may be the average value q_{2avg} , as shown in:

$$q_{avg}^2 = \frac{q_1 + q_2 + \dots + q_s + q + \Delta q}{s+1} = q_1^a + \frac{\Delta q}{s+1}. \quad (30)$$

From [Eqs. 28-30](#), the vectors $X_{S2 \cup PNZj}$ and X_{S2} exhibit significant dissimilarity because of the differing feature values $(q + \Delta q)$ and q_a . The only exception occurs when q_{a2} equals $q + \Delta q$ exactly. Therefore, the difference between $(X_{S2 \cup PNZj})$ and (X_{S2}) tends to be large. Consequently, the SHAP score for such dissimilar input features is given by:

$$\phi_{x_i^2}^2 = \frac{1}{Z} \left(\sum_{z=1}^Z \hat{f}(X_{S \cup PNZj}^{z2}) - \hat{f}(X_S^{z2}) \right), \quad (31)$$

It is generally high. From [Eqs. 27 and 31](#), it follows that.

$$\phi_{x_i^p}^2 \gg \phi_{x_i^p}^1. \quad (32)$$

Based on this observation, the second stage of the proposed algorithm leverages SHAP scores to differentiate between input vectors that are similar or dissimilar to the training set. This distinction helps determine whether the model's predictions can be trusted. It should be noted that a substantial difference in the absolute SHAP scores between the two scenarios occurs only when qa_2 does not take the value $q + \Delta q$. This limitation explains why the proposed Trust score may occasionally fail.

Trust Score: **Algorithm 1** presents the pseudo-code outlining the workflow of the proposed Trust Algorithm. With the forecast horizon denoted as n , the forecast step t starts at one and increases by one in each iteration until it reaches n . In every iteration, the algorithm compares the computed trust score T against the predefined Trust threshold T_{lim} . If the trust score is equal to or greater than the threshold, the prediction is considered reliable. On the other hand, if the trust score falls below the threshold, the forecast is likely to exhibit larger errors than the MAE or MAPE (depending on the evaluation metric used), indicating a poor or unreliable prediction. The threshold values ϕ_{lim} and T_{lim} are determined empirically based on observed prediction errors and their associated SHAP scores from the training and testing datasets.

Input: $T_c, T_{lim}, \phi_{max}, \phi_{sum}, \phi_{lim}$
Output: Trust score, T

- 1: **if** $(\phi_{sum} > \phi_{lim})$ **then**
- 2: $T = T_c / \phi_{sum}$
- 3: **else if** $(\phi_{max} > \phi_{lim})$ **then**
- 4: $T = T_c / \phi_{max}$
- 5: **else if** $(\phi_{sum} < \phi_{lim})$ and $\phi_{max} < \phi_{lim}$ and $T_c > T_{lim}$ **then**
- 6: $T = T_c \times \phi_{max}$
- OR
- 7: **else**
- 8: $T = T_c$
- 9: **end if**
- 10: **return** T

Algorithm 1. second stage for the proposed approach

Overall flow of the proposed trust algorithm

Algorithm 2 presents the pseudo-code for the complete workflow of the proposed approach. With the forecast horizon defined as n , the forecast step d starts at one and is incremented by one in each iteration until it reaches n . In every iteration, the algorithm evaluates the trust score T_d corresponding to each output element d and compares it against the Trust threshold T_{lim} .

If the trust score T_d is equal to or greater than the Trust threshold, the prediction is regarded as reliable. Conversely, when the trust score falls below the Trust threshold, the forecast is more likely to produce an error larger than the MAE or MAPE (depending on the specific error metric used for decision-making). In such cases, the prediction is considered unreliable or poor.

It should be noted that the performance of the algorithm is highly sensitive to the proper selection of the limiting parameters ϕ_{lim} , T_{limMAE} , and $T_{limMAPE}$.

Input: Forecast horizon n ,
 SHAP scores $\phi_{x_i^p}$ from same period previous year,
 PFI scores $\phi_{x_{EPF}}$, of previous year/training set,
 LR scores $\phi_{x_i^p}$,
 SHAP limit ϕ_{lim} decided based on model performance on training/test data,
 Trust Threshold value based on model performance on testing/training set for MAE T_{lim}^{MAE} and MAPE T_{lim}^{MAPE} , respectively,
 MAE_{*i*}/MAPE_{*i*} obtained from test set.

Output: Y_i^{MAE}/Y_i^{MAPE}

```

1: for  $d = 1$  to  $n$  do
2:   Stage 1: Obtain  $T_c$ .
3:   Stage 2: Obtain  $T$  using  $T_c$  of Stage 1
4:   if  $(T_d \leq T_{lim}^{MAE})$  then
5:      $Y_d^{MAE} = 1$ 
6:   else
7:      $Y_d^{MAE} = 0$ 
8:   end if
9:   OR
10:  if  $(T_d \leq T_{lim}^{MAPE})$  then
11:     $Y_d^{MAPE} = 1$ 
12:  else
13:     $Y_d^{MAPE} = 0$ 
14:  end if
15: end for
16: return  $Y_i^{MAE}/Y_i^{MAPE}$ 

```

Algorithm 2. Base algorithm for the proposed approach.

Results and Discussion

This section assesses the performance of the approach. We examine how effectively the proposed Trust algorithm can assess the reliability of forecasts generated by a convolutional neural network (CNN) for electricity price prediction (EPF). The evaluation relies on real datasets from the Italian power market and highlights the algorithm's ability to detect shifts in data distribution, known as data drift.

Subsection "Case Study 2: ERCOT Market" extends the application of this algorithm to a long short-term memory (LSTM) forecasting model trained on ERCOT market data. The findings confirm that the method remains robust and performs reliably even under conditions of substantially greater price fluctuations.

Case study 1: Italy's electricity market

The Italian power market includes both day-ahead spot trading and forward contracts [32]. Geographically, it is segmented into six distinct bidding areas: North, Central North, Central South, South, Sicily, and Sardinia. Wholesale prices from adjacent foreign markets, such as France, Austria, Greece, Corsica, and Slovenia, also shape market prices in these zones.

Buyers and sellers operating in each zone must deliver their offers and bids no later than the day before the trading window closes. The Gestore dei Mercati Energetici (GME), acting as Italy's independent system operator (ISO) for the day-ahead segment, establishes zonal clearing prices for generators and a uniform national reference price—referred to

as the “Prezzo Unico Nazionale” (PUN)—for purchasers. These values are derived through a merit-based ranking of offers, while also respecting the physical limitations of the transmission grid. The PUN is computed as a weighted mean that blends the various zonal prices and incorporates neighboring market prices only when GME authorizes cross-border energy transfers.

Detailed hourly records covering the PUN, individual zonal prices, and rates from neighboring countries were retrieved directly from the market operator’s official online platform [33].

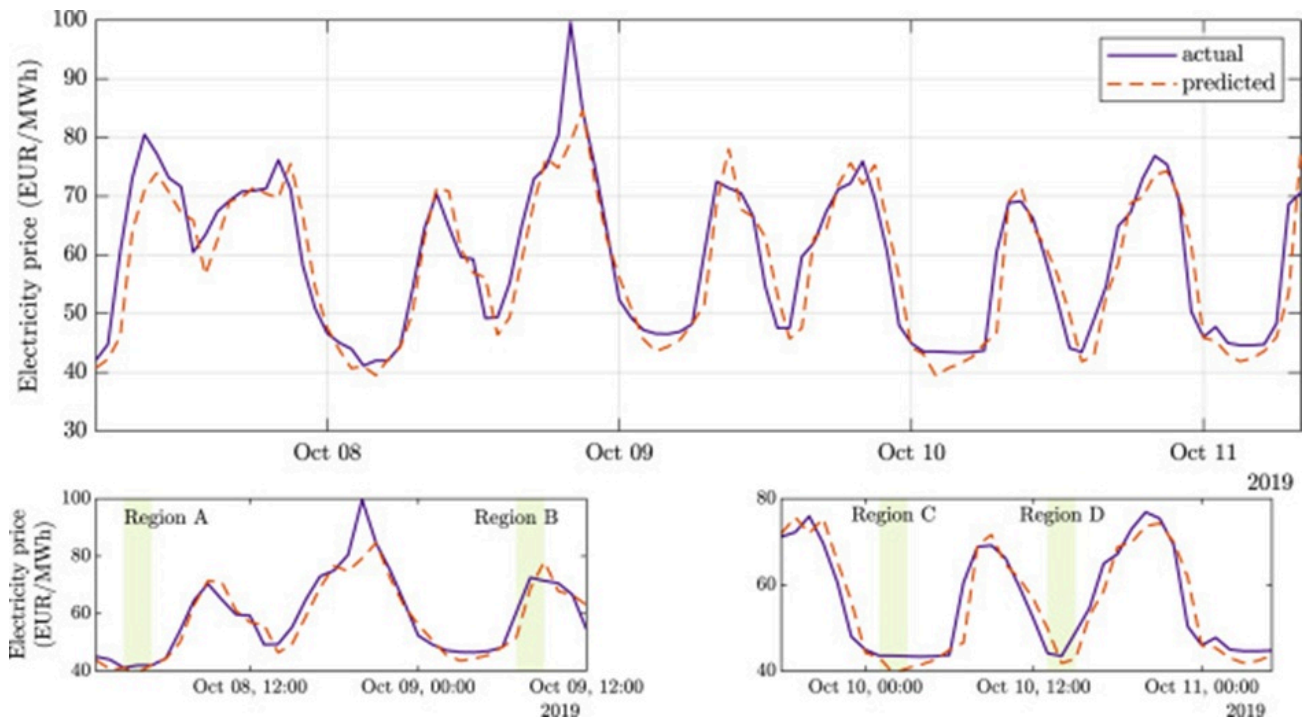


Figure 2. Forecasted hourly prices versus actual values for 100 successive observations covering the period from October 7 to October 11, 2019.

Explaining prediction reliability to electricity price forecasting users via the Trust algorithm

In this study, a convolutional neural network (CNN) model serves as the electricity price forecasting (EPF) tool to estimate Italy’s single national electricity price, known as PUN. The model takes as inputs the hour of the day, the daily average price, and the PUN values from the three previous hours (PUN at $t-2$, $t-1$, and t). In addition, it incorporates the previous hour’s electricity prices from the Central North (CNOR), Central South (CSUD), and Sicily (SICI) zones within Italy, along with prices from the neighboring virtual zones of Austria (AUST), Corsica (COAC), and Greece (GREC).

CNOR, CSUD, and SICI represent actual physical bidding zones of the Italian electricity grid, whereas AUST, COAC, and GREC correspond to virtual foreign zones whose prices influence the Italian market. These input variables were chosen based on the observed correlation between prices in Italy’s domestic physical zones and those in adjacent foreign virtual zones. Highly correlated features were eliminated, retaining only one representative feature to capture their collective effect on the PUN.

The CNN model was trained on hourly data from the Italian electricity market for the years 2015, 2016, and 2017, with 2018 data reserved for testing. Validation of the proposed Trust algorithm was performed using market data from 2019, 2020, and 2021.

Feature-importance explanations for accurate and inaccurate forecasts

This part of the analysis focuses on a limited segment of the 2019 predictions—specifically, a sequence of one hundred consecutive hourly forecasts spanning October 7 to October 11, 2019. By examining this set of predictions and their associated explanations, the subsection demonstrates why conventional explanation techniques offer limited practical value to end users of electricity price forecasts.

Figure 2 presents a comparison between the CNN model's predicted prices and the realized electricity prices over the selected period. The mean absolute error (MAE) across this region stands at 2.716 EUR/MWh. As illustrated in **Figure 3**, the prediction error fluctuates around this MAE value for every individual forecast. Out of the 100 predictions, exactly 50 exhibit errors exceeding the MAE, while the remaining 50 show smaller deviations.

In practical trading environments, the MAE for a given forecasting window is known only after the actual market prices are observed in real time. Consequently, when users need to forecast hour-ahead prices for 2019, they have access solely to the model's MAE from the previous year (2018), which was 2.946 EUR/MWh. Using this 2018 MAE as a benchmark, 47 of the 100 predictions in the selected period display errors larger than the reference value. The maximum absolute error observed reaches 20.38 EUR/MWh—nearly seven times the reference MAE. **Figure 3** directly contrasts the individual prediction errors against this 2018 benchmark, revealing that summary error statistics fail to provide meaningful guidance at the level of single data points.

First, the absolute error surpasses the MAE in numerous cases. Second, it remains unclear in advance when such large deviations will occur. As a result, electricity price forecasting users cannot determine when the risk associated with a particular prediction exceeds acceptable levels or when a forecast is sufficiently reliable to support profitable bidding decisions.

SHAP-based explanations are primarily useful for AI specialists, as they reveal how the model reacts to various input features. This allows experts to assess whether the model accurately reflects the market's underlying stochastic behavior and to refine the architecture for greater accuracy. However, the requirements of EPF end-users differ substantially. These users seek explanations that deliver actionable insight into the likely quality or trustworthiness of each prediction, enabling them to adjust their bidding strategies accordingly and make optimal use of the forecast.

To evaluate whether SHAP visualizations meet the practical needs of EPF users, four groups of consecutive data points were selected from the 100-point sequence. Two groups correspond to relatively accurate predictions (errors below the MAE), while the other two represent poor predictions (errors above the MAE). Details of these selected points are provided in **Table 1**, with their locations indicated graphically in **Figure 2**. Enlarged views of these four regions appear in the right-side and bottom-left subplots of **Figure 2**.

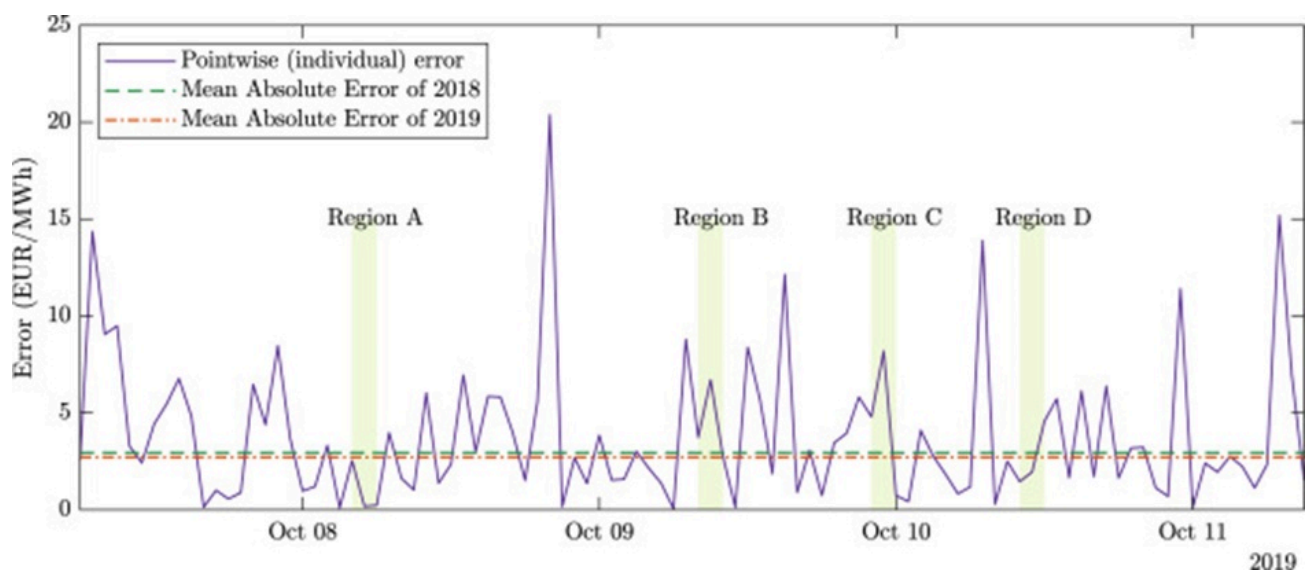


Figure 3. Comparison between individual prediction errors and the mean absolute error (MAE) values from both 2018 and 2019 across the hourly data points spanning 6:00 on October 7 to 09:00 on October 11.

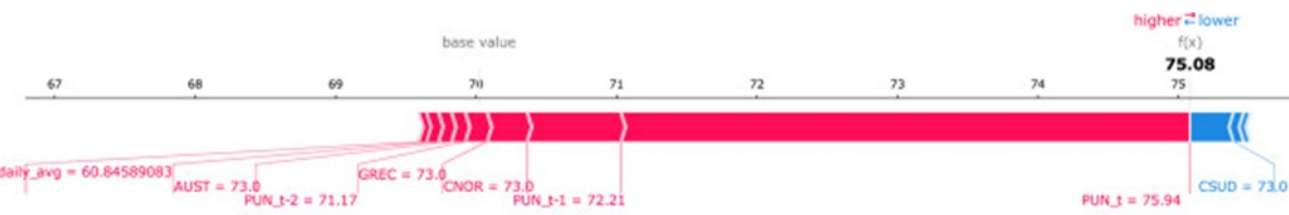


Figure 4. Force ploregion C, which has bad predictions.



Figure 5. Force plot of region D, which has good predictions.

From the presented table and figures, it is evident that the predictions in regions B and C are unreliable, as their errors exceed the reference mean absolute error (MAE) of 2.83 EUR/MWh. In contrast, the forecasts in regions A and D prove more accurate, with errors remaining below this reference threshold. The feature-level explanations for these four regions are examined next.

Figures. 4 and 5 display the force plots, while **Figures. 6 and 7** present the waterfall plots for data points from regions with poor (region C) and strong (region D) performance. These visualizations present SHAP-based interpretations of the machine learning model's outputs in the selected areas. They illustrate how the influence of each input feature varies across successive predictions.

Although the plots reveal noticeable differences in feature contributions between accurate and inaccurate forecasts, no consistent pattern or clear rule emerges from these visualizations that would allow electricity price forecasting (EPF) users to reliably determine whether a specific prediction is likely to be “good” or “bad.”

This limitation does not imply that SHAP values lack useful content. In reality, the SHAP values encode meaningful signals about the reliability and quality of each model prediction. However, this critical information remains largely inaccessible to end-users in both visual and numerical forms.

The proposed Trust algorithm addresses this gap by enabling EPF users to access such insights through a straightforward, score-based explanation. The algorithm constructs a trust score by analyzing the intrinsic properties of the explanation values and their relationships.

Table 1. Good and bad predictions of different forecast regions.

Region	Type	Actual	Predicted	Error
A	Good	41.14	41.045	−0.095
Empty Cell		42.00	39.467	−2.533

Empty Cell		42.00	42.187	−0.187
B	Bad	60.46	51.663	−8.797
Empty Cell		72.51	68.772	−3.737
Empty Cell		71.40	78.106	6.706
C	Bad	69.58	75.396	5.816
Empty Cell		60.63	65.434	4.804
Empty Cell		47.99	56.181	8.191
D	Good	68.86	69.158	0.298
Empty Cell		69.15	71.652	2.502
Empty Cell		66.19	64.731	−1.459

Score-based explanation provided by the proposed Trust algorithm for reliable and unreliable predictions

This subsection revisits the same forecast segments analyzed earlier. It demonstrates how the trust score produced by the suggested algorithm effectively addresses the practical explanation requirements of electricity price forecasting (EPF) users.

The Trust algorithm relies on several inputs, including SHAP values, permutation feature importance (PFI) scores, linear regression (LR) coefficients, and error metrics such as MAE and MAPE. Selected portions of these supporting data are summarized in [Tables 2 and 3](#).

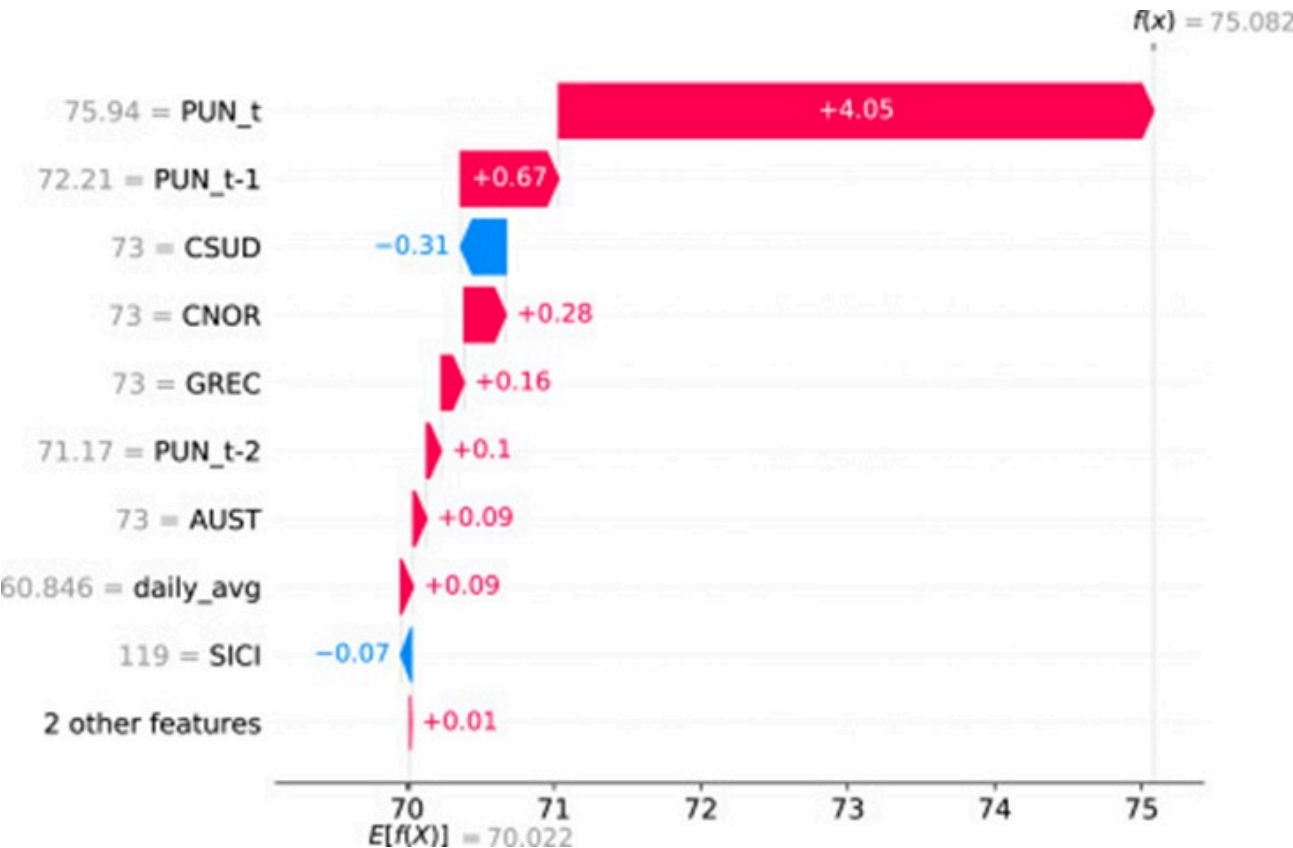


Figure 6. Waterfall plot for region C, which shows poor predictions.

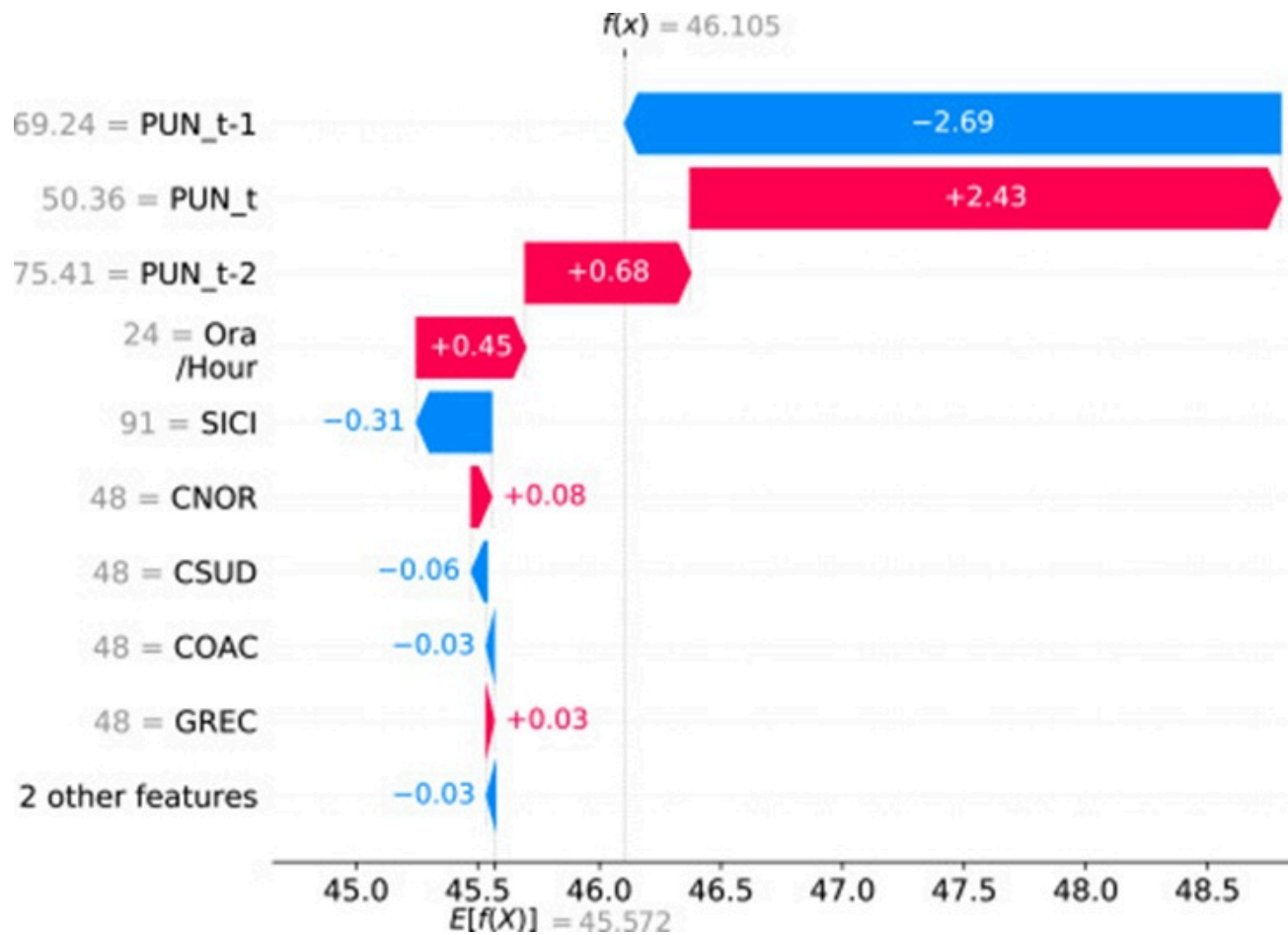


Figure 7. Waterfall plot of region D, which has good predictions.

Table 2 summarizes the various error indicators calculated for the 2018 period. In Table 3, one can find the permutation feature importance scores extracted from the 2018 dataset, together with the linear regression coefficients fitted on the training data spanning 2015 through 2017.

In the current case study, the Trust algorithm applies a threshold value of 0.75 as the acceptable limit for both MAE and MAPE, referred to here as the Trust Threshold (TMAE/MAPElim). These boundary values were established after careful examination of the training years (2015, 2016, and 2017) and the test year (2018). It should be emphasized that such thresholds are not universal; they will shift whenever a different forecasting model is employed or when the underlying dataset changes.

Likewise, the acceptable SHAP limit, denoted ϕ_{lim} , was set based on insights from the preceding year’s data. In this analysis of the Italian electricity market, ϕ_{lim} was set to 6 for both the MAE and MAPE calculations.

Table 2. Error metrics for 2018 are available to the end user.

Sr. no.	Value	Error metric
1	0.92	R2 score
2	5.179	Mean Percentage Error

3	2.946	Mean Absolute Error
---	-------	---------------------

Table 4 illustrates the behavior of the proposed algorithm across the four selected prediction segments (regions A, B, C, and D). As shown in **Table 4**, the algorithm consistently assigns a trust score higher than 0.75 to every prediction whose individual error stays below the mean absolute error (MAE). Conversely, it assigns a trust score below 0.75 to all predictions where the individual error exceeds the overall MAE.

The algorithm operates in two stages. In the initial stage, it generates a preliminary trust score based on correlation coefficients ranging from -1 to 1 . The subsequent fine-tuning stage then shifts this score closer to either 0 or 1.

From **Table 4**, it is clear that in the regions with poor predictions, the resulting trust scores tend to cluster near 0.1, whereas in the regions with accurate predictions, the scores approach 0.9.

Electricity price forecasting (EPF) users can leverage these trust scores to manage risk better when developing bidding strategies or making scheduling decisions. For instance, when the trust score falls below 0.75, bidding optimization tools can be configured to adopt a more conservative, risk-averse approach that accounts for the elevated chance of significant forecasting errors.

Alternatively, users may choose to disregard the machine learning prediction entirely and rely on the judgment of experienced asset managers, who can apply their domain knowledge to make the best decision. A third possibility involves switching to other forecasting methods.

In all these scenarios, asset managers gradually develop greater confidence in the ML model because they gain clear insight into when its predictions are likely to be reliable and when they are not.

Table 3. PFI score and LR coefficients of the CNN-based EPF model.

Sr. no.	Feature name	LR coefficients	PFI score
1	Ora/Hour	0.09953	0.01664
2	Daily avg	0.22786	0.02006
3	PUN $t-2$	-0.13863	0.04334
4	PUN $t-1$	-0.29255	0.04557
5	PUN t	1.01141	0.21793
6	AUST	0.14106	0.02410
7	CNOR	-0.00115	0.02339
8	COAC	0.00355	0.00454
9	CSUD	0.00534	0.00666
10	GREC	0.01021	0.00849
11	SICI	0.00302	0.00213

Table 4. Predictions with Trust Scores with a Trust Threshold of 0.75.

Region	Type	Actual	Predicted	Error	Trust score
--------	------	--------	-----------	-------	-------------

A	Good	70.79	72.87	2.08	0.8466
Empty Cell		70.95	72.79	1.84	0.9837
Empty Cell		71.33	71.91	0.58	0.9581
B	Bad	71.2	78.88	7.68	0.0848
Empty Cell		58.27	68.83	10.56	0.0855
Empty Cell		50.99	54.80	3.81	0.151
C	Bad	69.58	78.34	8.76	0.09215
Empty Cell		60.63	66.79	6.16	0.09305
Empty Cell		47.99	57.63	9.64	0.1119
D	Good	44.96	44.49	−0.469	0.9163
Empty Cell		43.55	44.15	0.599	0.8736
Empty Cell		43.55	40.72	−2.82	0.9907

Model performance

To assess how effectively the proposed algorithm generates accurate trust scores, two correctness metrics are introduced for good and bad predictions:

$$C_{good} = \frac{T_{good}}{T_{good} + F_{bad}} \times 100\% \quad (33)$$

$$C_{bad} = \frac{T_{bad}}{T_{bad} + F_{good}} \times 100\%. \quad (34)$$

Here, C_{good} denotes the percentage of actual good predictions that the Trust algorithm correctly identifies (out of all truly good predictions, including those mistakenly classified as bad). Similarly, C_{bad} represents the percentage of actual bad predictions that are properly recognized. In these definitions, T_{good} denotes the number of correctly identified good predictions, while F_{bad} denotes the number of correctly identified good predictions incorrectly labeled as bad. On the other hand, T_{bad} and F_{good} indicate the number of correctly and incorrectly classified bad predictions, respectively.

Additionally, to evaluate the algorithm's overall classification performance, two further indicators are defined as follows:

$$I_{corr} = \frac{T_{good} + T_{bad}}{T_{good} + T_{bad} + F_{good} + F_{bad}} \times 100\%, \quad (35)$$

$$I_{incorr} = \frac{F_{good} + F_{bad}}{T_{good} + T_{bad} + F_{good} + F_{bad}} \times 100\%. \quad (36)$$

I_{corr} measures the algorithm's ability to correctly distinguish between good and bad predictions, whereas I_{incorr} captures the rate of misclassifications (good predictions labeled as bad, and bad predictions labeled as good).

These metrics demonstrate the model's effectiveness when applied to data from 2020 and 2021.

Model performance for the year 2020

Tables 5 and 6 report the values of C_{good} , C_{bad} , I_{corr} , and I_{incorr} obtained by evaluating the Trust algorithm's capability to assess the quality of predictions generated by the CNN model on 2020 data. Results in **Table 5** are based on the mean absolute error (MAE), while those in **Table 6** rely on the mean absolute percentage error (MAPE).

The tables were constructed using 10 randomly chosen clusters, each consisting of 100 consecutive hourly data points. The results indicate that the algorithm performs well at identifying good predictions. However, it appears less effective at detecting bad predictions (This observation is examined in greater depth in the following subsection using 2021 data, which reveals that the algorithm is indeed capable of identifying bad predictions. The lower performance occurs primarily when dealing with the minority class of predictions, regardless of whether they are good or bad).

Across the 1000 data points considered in the MAE and MAPE tables, the total number of correctly classified predictions stands at 739 and 617, respectively. Overall, the algorithm demonstrates greater effectiveness when prediction quality is evaluated using absolute error (MAE) compared to percentage-based error (MAPE).

Table 5. Performance with respect to $YMAE$ for different time periods of 2020.

Time range (from/to)		<i>C_{good}</i>	<i>C_{bad}</i>	<i>I_{corr}</i>	<i>I_{incorr}</i>
1 Jan, 01:00	5 Jan, 04:00	88.46	27.27	75	25
3 Mar, 12:00	7 Mar, 16:00	90.00	46.67	77	23
1 Apr, 17:00	5 Apr, 21:00	85.71	32.43	66	34
14 Apr, 05:00	18 Apr, 09:00	94.2	29.03	74	26
25 May, 21:00	30 May, 01:00	90.0	20.0	76	24
06 Jul, 13:00	10 Jul, 17:00	93.9	27.78	82	18
17 Aug, 05:00	21 Aug, 09:00	91.89	19.2	73	27
27 Sep, 21:00	2 Oct, 01:00	88.73	51.7	78	22
8 Nov, 12:00	12 Nov, 16:00	81.33	40.00	71	29
20 Dec, 04:00	24 Dec, 08:00	86.56	27.27	67	33

Table 6. Performance with respect to $YMAPE$ for different time periods of 2020.

Time range (from/to)		<i>C_{good}</i>	<i>C_{bad}</i>	<i>I_{corr}</i>	<i>I_{incorr}</i>
1 Jan, 01:00	5 Jan, 04:00	90.00	22.50	63	37
3 Mar, 12:00	7 Mar, 16:00	89.36	30.18	58	42
1 Apr, 17:00	5 Apr, 21:00	91.17	27.27	49	51
14 Apr, 05:00	18 Apr, 09:00	95.35	19.29	52	48
25 May, 21:00	30 May, 01:00	90.24	13.56	45	55
06 Jul, 13:00	10 Jul, 17:00	94.03	18.18	69	31
17 Aug, 05:00	21 Aug, 09:00	94.91	19.51	64	36
27 Sep, 21:00	2 Oct, 01:00	90.16	43.58	72	28

8 Nov, 12:00	12 Nov, 16:00	86.56	45.45	73	27
20 Dec, 04:00	24 Dec, 08:00	89/70	34.37	72	28

The confusion matrix for the 1000 data points analyzed in [Table 5](#) is shown in [Figure 8](#). The top-left section of the figure highlights the correct and incorrect classifications of accurate versus inaccurate predictions. The top-right confusion matrix displays the misclassification percentages from the perspective of the actual good and bad predictions, which are known only after the market observes the actual electricity prices. In contrast, the bottom-left confusion matrix reflects the perspective of electricity price forecasting (EPF) users.

Model oerformance for the Year 2021

Electricity prices throughout 2020 closely resembled the patterns the CNN model had encountered during its training phase (2015–2017). In 2021, however, a sharp increase in natural gas prices triggered an extraordinary surge in electricity prices starting around mid-year [\[34\]](#). [Figure 9](#) compares the actual electricity prices observed in 2021 with those from 2015 to 2020, clearly showing that both the scale and the behavior of prices in the second half of 2021 deviate substantially from the data used to train the model. As a result, the CNN model is forced to extrapolate when handling the out-of-distribution (OOD) segment of 2021, leading to a noticeable decline in forecasting accuracy. This deterioration is evident in [Figure 10](#). In the early months of 2021, the CNN model's performance remains comparable to that in 2018, 2019, and 2020. However, in the latter part of 2021, the prediction errors increased dramatically compared to the same period in previous years.

This situation exposes EPF users to substantial financial risks because the model's forecasts are unreliable during periods of rapid price escalation. Consequently, the ability to detect such shifts in data distribution becomes critically important.

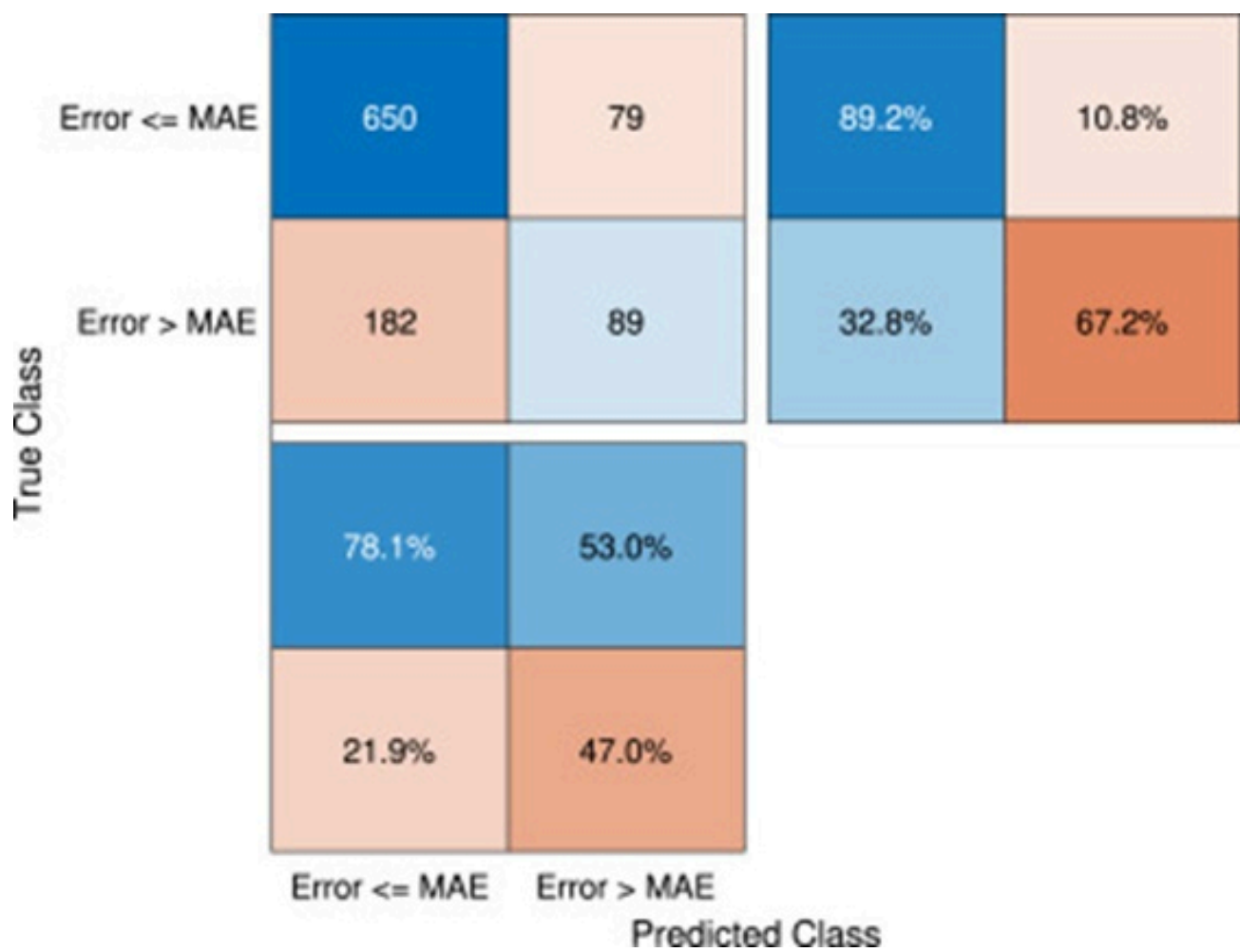


Figure 8. Confusion matrix for the 1000 data points discussed in Table 5, highlighting the overall performance of the model on the 2020 data.

The performance of the proposed algorithm when applied to the 2021 dataset is summarized in Table 7. Figure 11 is constructed using the correctly identified severe (poor) predictions—specifically the True Negative cases—for the year 2021. When these values are compared against the True Negative counts from 2020, the results demonstrate that the Trust algorithm successfully detects the shift in data distribution.

Such an alert generated by the Trust algorithm enables electricity price forecasting (EPF) users to implement appropriate protective actions in advance. One effective response could involve consulting the AI development team to refine or even replace the forecasting model as needed.

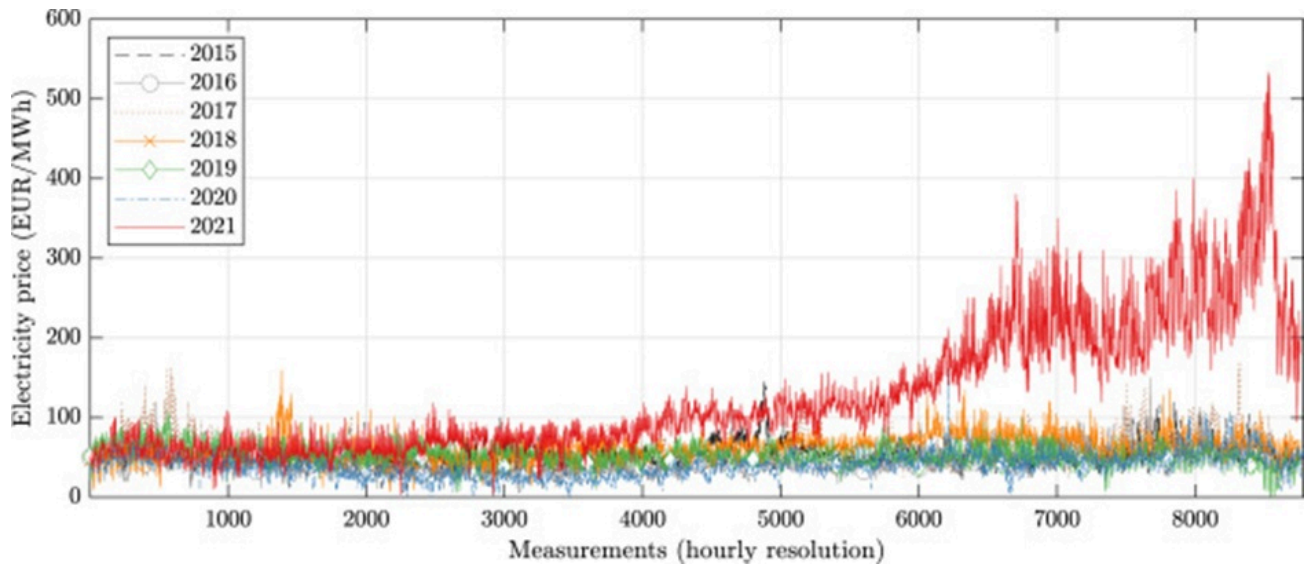


Figure 9. Data from 2019 and 2020 is IID, while that of 2021 is OOD.

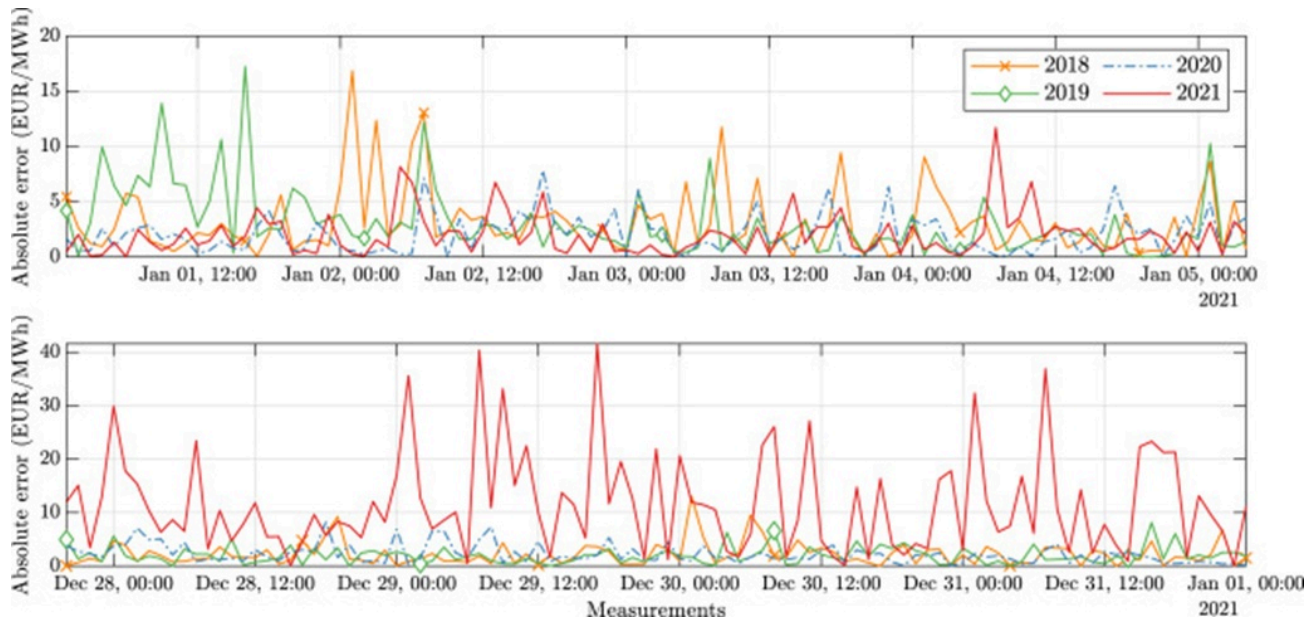


Figure 10. Absolute errors seen during the first and last 100 consecutive points for the years 2018, 2019, 2020, and 2021.

The confusion matrix for the 2021 data is shown in **Figure 12** (alongside the previous matrix for direct comparison). For the initial 500 data points—where the 2021 observations remain in-distribution (IID) relative to the training data—the proposed Trust algorithm correctly classifies $256 + 103 = 359$ of the CNN model's predictions. From the electricity price forecasting (EPF) user's perspective, this corresponds to 74.2% of good predictions and 66.5% of bad predictions being correctly identified. The corresponding C_{good} and C_{bad} values for this segment are 83.1% and 53.6%, respectively.

In contrast, the final 500 data points of 2021 fall into the out-of-distribution (OOD) region due to the data shift. Here, the algorithm correctly identifies $76 + 254$ predictions. Notably, 80.9% of the actual bad predictions are successfully flagged by the algorithm, resulting in a substantially higher C_{bad} percentage than typically observed.

To provide an overall view of the algorithm's performance across the full year 2021, the confusion matrix covering all 1000 data points referenced in **Table 7** is presented in **Figure 13**.

In summary, although the CNN forecasting model's accuracy declines noticeably when moving from IID to OOD data, the proposed Trust algorithm continues to deliver consistent performance in distinguishing between reliable and unreliable machine learning predictions.

Table 7. Performance with respect to YMAE for different time periods of 2021.

Time range (from/to)		<i>Cgood</i>	<i>Cbad</i>	<i>Icorr</i>	<i>lincorr</i>	TN	TP	FN	FP
1 Jan, 01:00	5 Jan, 04:00	92.4	33.3	80	20	7	73	6	14
4 Mar, 12:00	8 Mar, 16:00	85.9	58.6	78	22	17	61	10	12
2 Apr, 17:00	6 Apr, 21:00	74.0	48.0	61	39	24	37	13	26
15 Apr, 05:00	19 Apr, 09:00	77.2	51.2	66	34	22	44	13	21
26 May, 21:00	31 May, 01:00	80.4	67.3	74	26	33	41	10	16
7 Jul, 13:00	11 Jul, 17:00	79.3	44.7	63	37	21	42	11	26
18 Aug, 05:00	22 Aug, 09:00	66.7	39.1	49	51	25	24	12	39
28 Sep, 21:00	3 Oct, 01:00	10.0	73.8	61	39	59	2	18	21
9 Nov, 12:00	13 Nov, 16:00	33.3	79.3	71	29	65	6	12	17
21 Dec, 04:00	25 Dec, 08:00	22.2	92.3	86	14	84	2	7	7

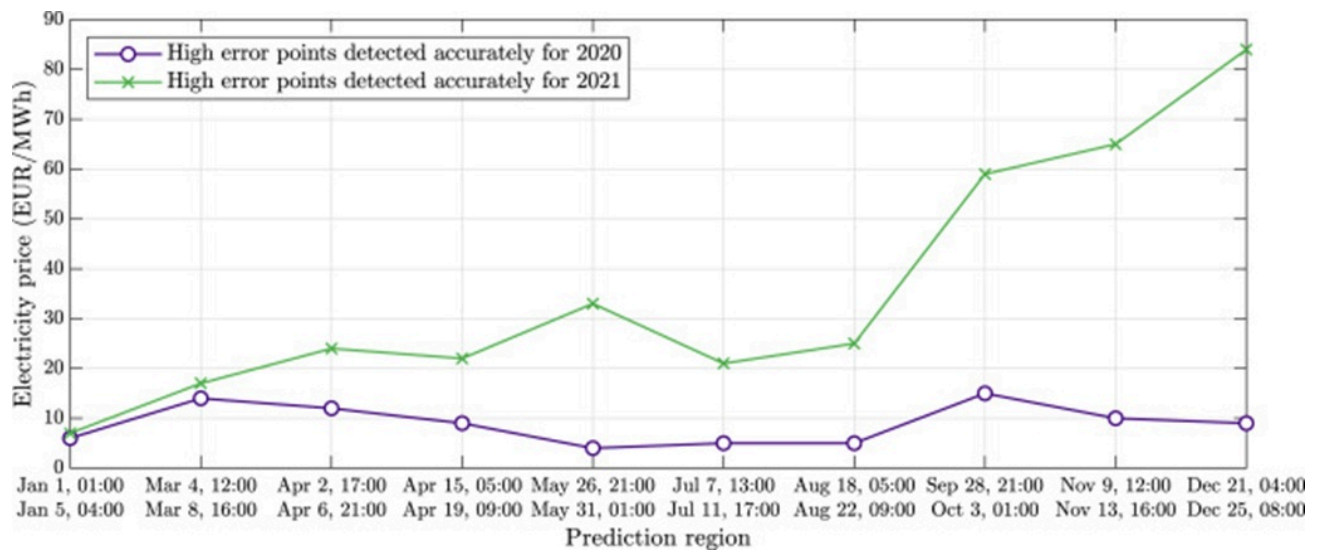
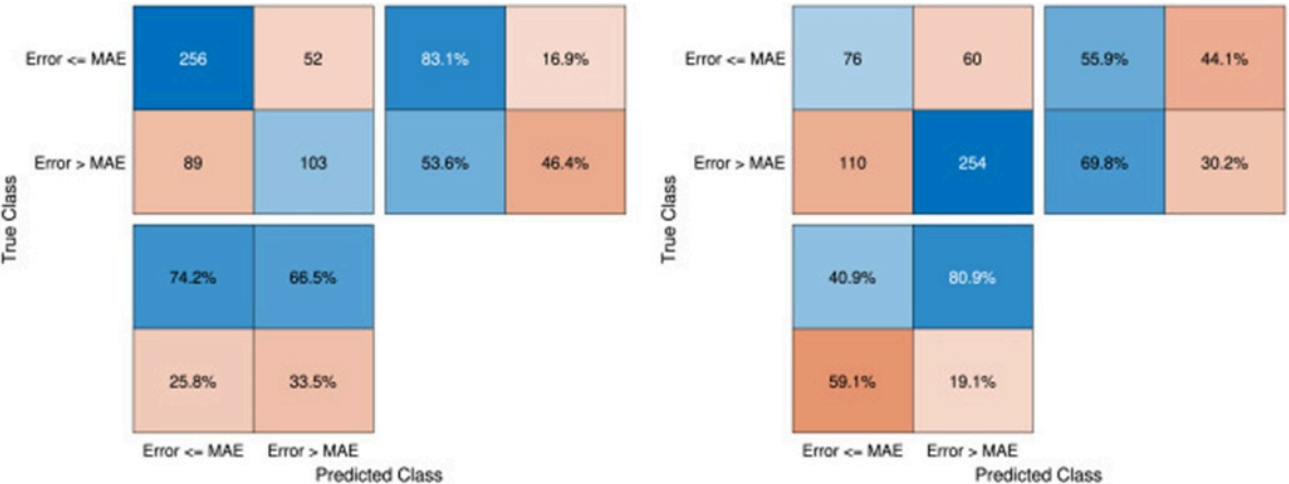


Figure 11. Distributional data shift is reflected in the Trust algorithm.



(a) Confusion matrix for the first 500 data points from Table 7 which represent IID region of 2021

(b) Confusion matrix for the last 500 data points from Table 7 which represent OOD region of 2021

Figure 12. Confusion matrix for 2021 split into two half-periods.

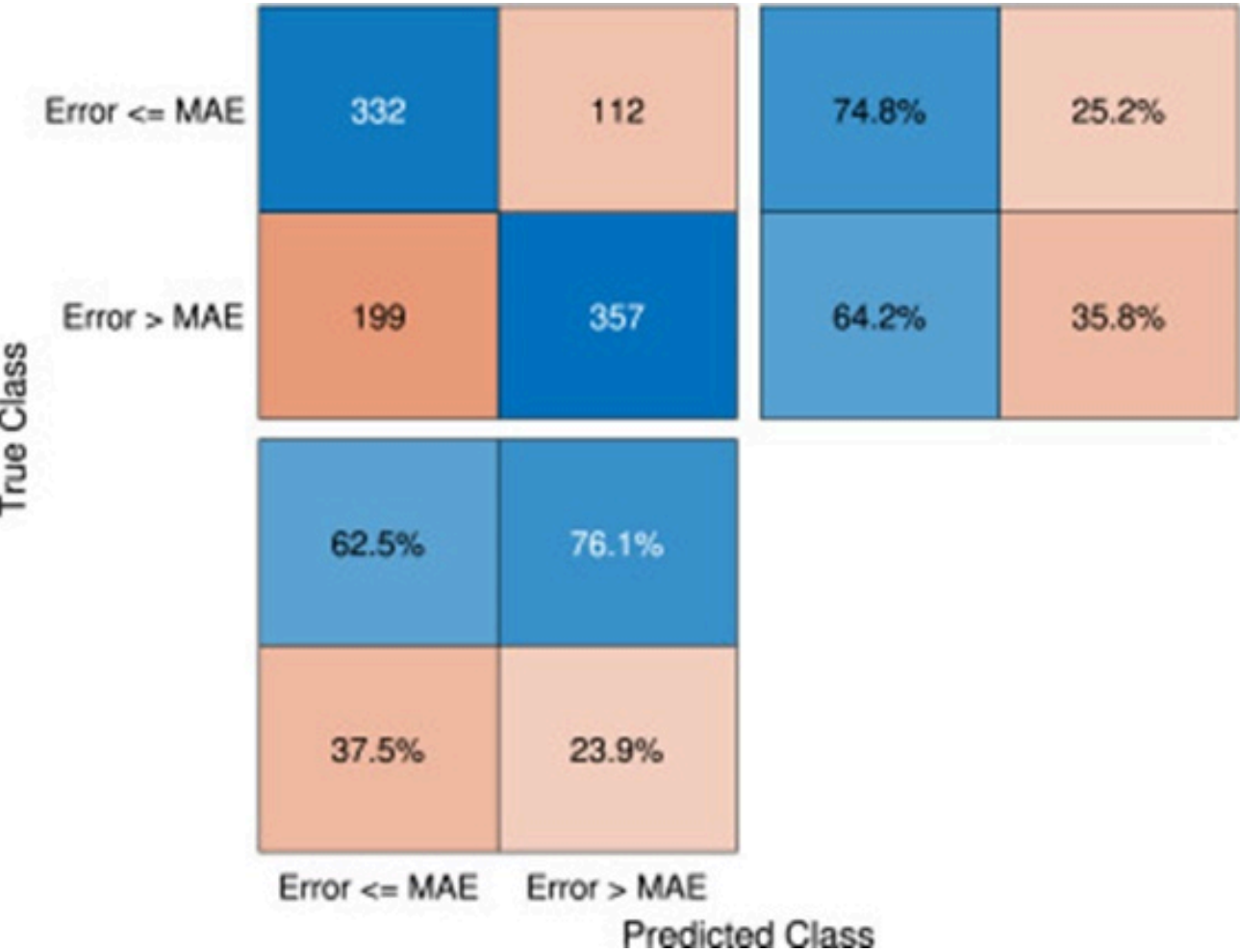


Figure 13. The overall confusion matrix for the 1000 data points mentioned in Table 7 illustrates the model's performance in 2021.

Case study 2: ERCOT market

The Electric Reliability Council of Texas (ERCOT) functions as a major regional reliability organization within the North American Electric Reliability Corporation framework. It manages both power grid operations and wholesale electricity markets for the ERCOT area, which covers nearly 85% of the state of Texas [35].

A key feature that distinguishes the ERCOT market from the Italian one is the frequent occurrence of extreme price spikes, which began after the introduction of the nodal market design and the day-ahead market in 2010 [36]. These sharp price surges typically occur during high-demand periods in both winter and summer.

For instance, prices in the Houston zone have fluctuated widely, reaching as low as 5.1 USD/MWh and as high as 2060.4 USD/MWh in 2018, and ranging from 3.8 USD/MWh to a peak of 5011.6 USD/MWh in 2019 [37]. Leaving out these extreme spikes from training data often leads machine learning models to break down during peak events, while incorporating them tends to degrade the model's general performance. The proposed Trust algorithm helps electricity price forecasting (EPF) users distinguish between predictions that can be trusted and those that should be rejected or verified using other forecasting tools or expert judgment.

Providing reliability insights to EPF users through the trust algorithm

To examine how well the proposed algorithm performs in evaluating forecast quality, an LSTM-based electricity price forecasting model was developed specifically for one ERCOT zone.

The ERCOT territory is organized into multiple bidding zones, such as North (LZ_NORTH), South (LZ_SOUTH), West (LZ_WEST), Rayburn (LZ_RAYBN), and Houston (PRED_HOUSTON) [36]. The South zone itself is divided into smaller areas, including Austin Energy (LZ_AEN), Lower Colorado River Authority (LZ_LCRA), and CPS Energy (LZ_CPS) [36].

The LSTM model takes several inputs: the hour of the day (HOUR), overall system load (LOAD), prices from neighboring zones, and recent historical prices from the target zone. In this analysis, the focus is on forecasting the Houston zonal price at time t . Accordingly, the model includes the Houston prices from the three preceding hours ($t-1$, $t-2$, and $t-3$) as input variables. In addition, prices for three ancillary services—up-regulation (UP REG), responsive reserve service (RRS), and non-spin service (NON SPIN)—are factored in.

The model was trained using hourly records from 2015, 2016, and 2017, and subsequently tested on 2018 data. Evaluation of the Trust algorithm's effectiveness on ERCOT data is covered in the next two subsections.

Performance evaluation on 2019 ERCOT data

One of the defining traits of the ERCOT market is the frequent occurrence of abrupt, intense price spikes. While the LSTM model generally delivers acceptable results for most of the year, its performance collapses during these spike episodes. This inconsistency can cause EPF users to lose faith in the model and manually override its suggestions, despite its solid performance during normal market conditions. A reliable trust score accompanying each forecast can help users decide precisely when to rely on the LSTM output and when to approach it with caution.

To highlight this advantage, two separate periods are analyzed. The first period, labeled region A, runs from 11:00 pm on January 21 to 02:00 am on January 26 — a timeframe during which the LSTM model achieves strong accuracy. The second period, labeled region B, extends from 12:00 am on September 4 to 03:00 am on September 8 and coincides with a major price spike. During region B, the LSTM model's forecasting accuracy declines sharply, as illustrated in **Figure 14**.

Region A consists of 100 consecutive hourly points, out of which the LSTM model generates 88 predictions with errors below the 2018 mean absolute error (MAE). In region B, also containing 100 points, only 66 predictions achieve this level of accuracy.

For region A, the Trust algorithm correctly classifies 87.5% of the good predictions and 75% of the bad ones. In region B, it correctly identifies 74.24% of the good predictions and successfully flags 100% of the severely inaccurate predictions. These outcomes are visually presented in **Figures 15 and 16**. Importantly, every single poor prediction during the price spike in region B was detected by the algorithm. This outcome occurs because both the highest individual SHAP value and the cumulative sum of SHAP values at those spike points are considerably larger than those observed elsewhere.

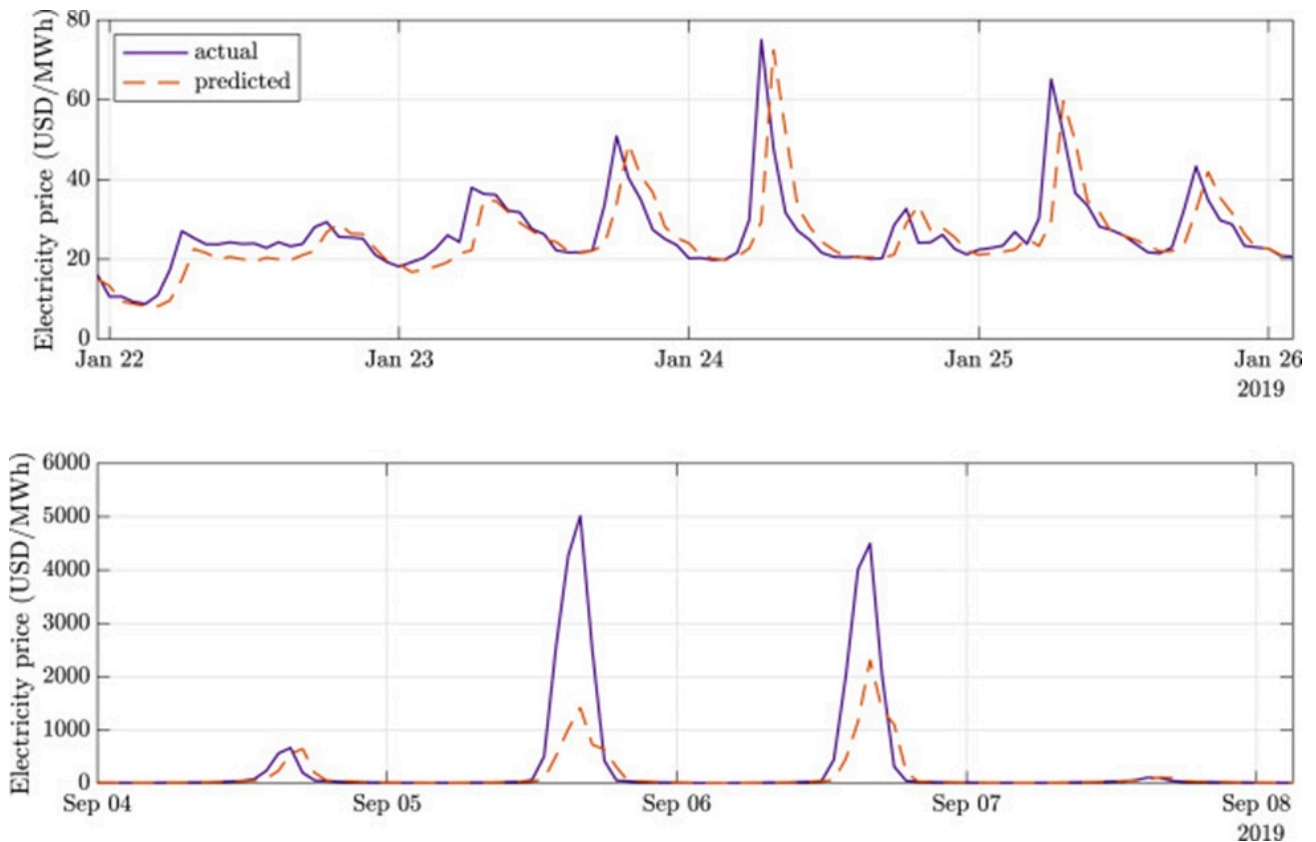


Figure 14. LSTM model forecasts versus actual prices for region A (from 11:00 pm on January 21 to 02:00 am on January 26) and region B (from 12:00 am on September 4 to 03:00 am on September 8).

Performance of the proposed model on 2021 ERCOT data

The surge in natural gas prices clearly influenced electricity prices in the ERCOT market during 2021, as in the Italian market. Nevertheless, the scale of price volatility observed in ERCOT that year was far more extreme than in Italy. In the Houston zone alone, prices fluctuated from a low of 0.06 USD/MWh to an extraordinary high of 8994.6 USD/MWh. Unlike the gradual upward trend in the Italian market, ERCOT experienced a sharp, abrupt price spike in February.

To assess the proposed Trust algorithm's capability under these challenging conditions, the LSTM model's predictions were examined across two distinct periods. The first period, designated as region C, covers the interval with the most intense price peaks in February. It includes 500 consecutive hourly predictions from 01:00 pm on February 5 to 08:00 am on February 26, as illustrated in **Figure 17a**. This segment corresponds to the portion where the LSTM model exhibits the largest prediction errors.

The second period consists of another set of 500 consecutive hourly predictions spanning from 8:00 pm on November 13 to 03:00 pm on December 4, as shown in **Figure 17b**. In this interval, the LSTM model maintains relatively strong performance.

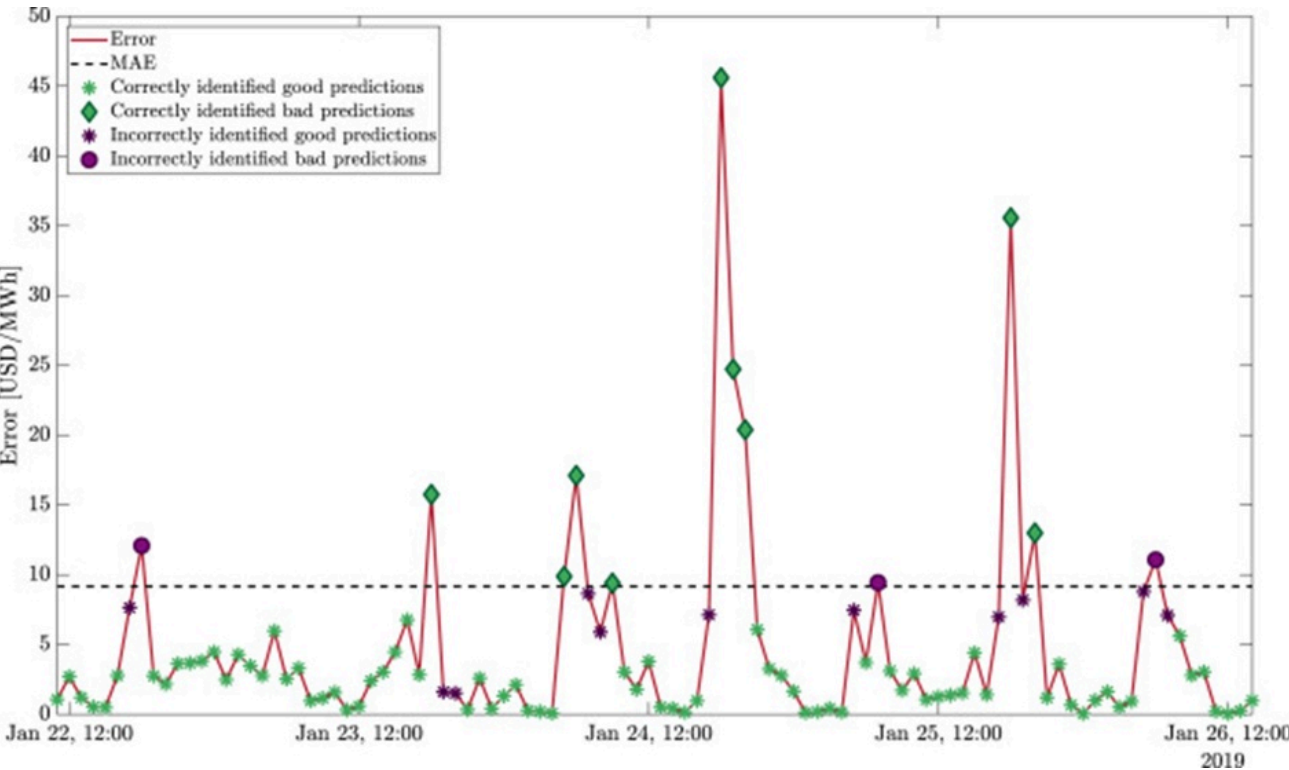


Figure 15. Evaluation of the proposed Trust algorithm's ability to assess the quality of LSTM forecasts in region A for the 2019 ERCOT market data.

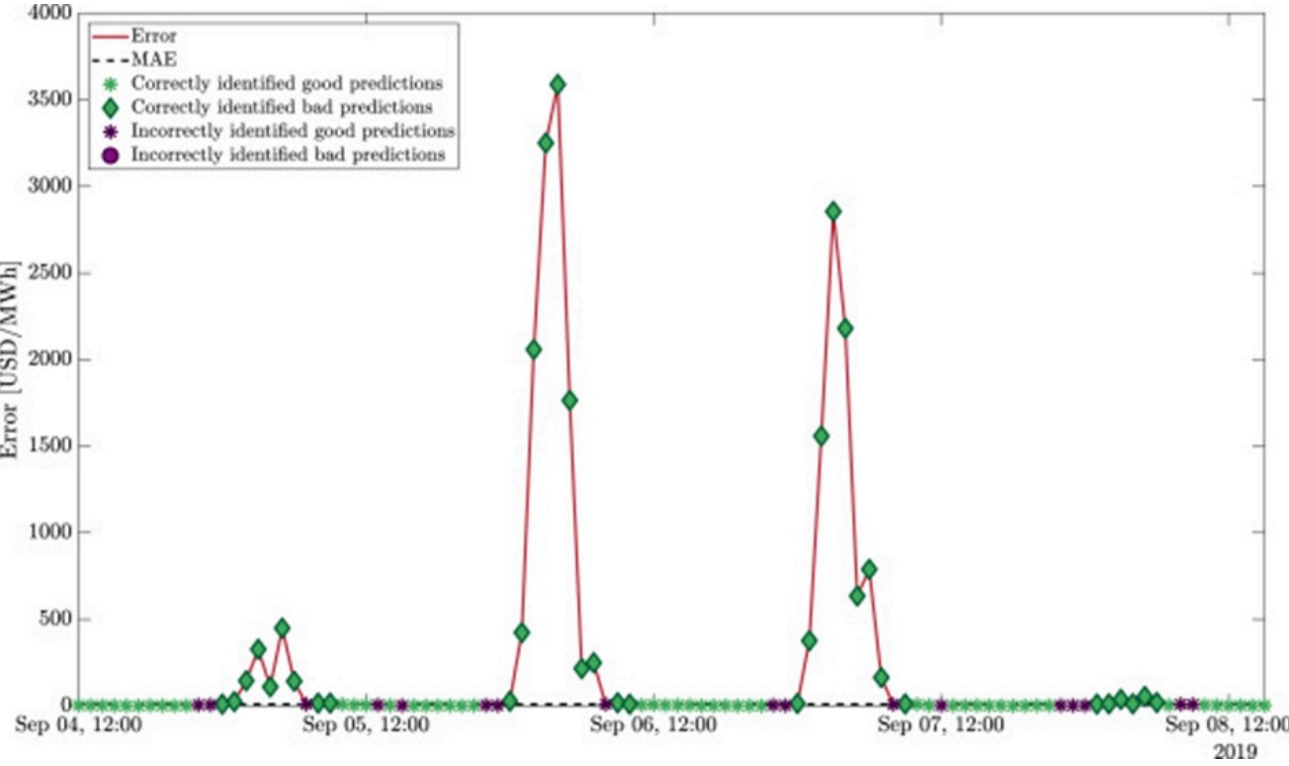


Figure 16. Assessment of the proposed Trust algorithm's effectiveness in evaluating the quality of LSTM model predictions within region B for the 2019 ERCOT market data.

Table 8 presents the performance of the proposed algorithm across region C, the most problematic segment of the year. The results indicate that the algorithm performs strongly in the central parts (sub-regions C3 and C4) as well as the later portions (sub-regions C1 and C6) of this challenging interval.

Table 9 summarizes the effectiveness of the Trust score algorithm in region D. The findings demonstrate that the proposed trust scoring mechanism successfully differentiates between reliable and unreliable predictions during this period, even though the training data from 2015 to 2017 contained neither the extreme price levels observed in 2021 nor the sharp increase in electricity prices driven by rising natural gas costs.

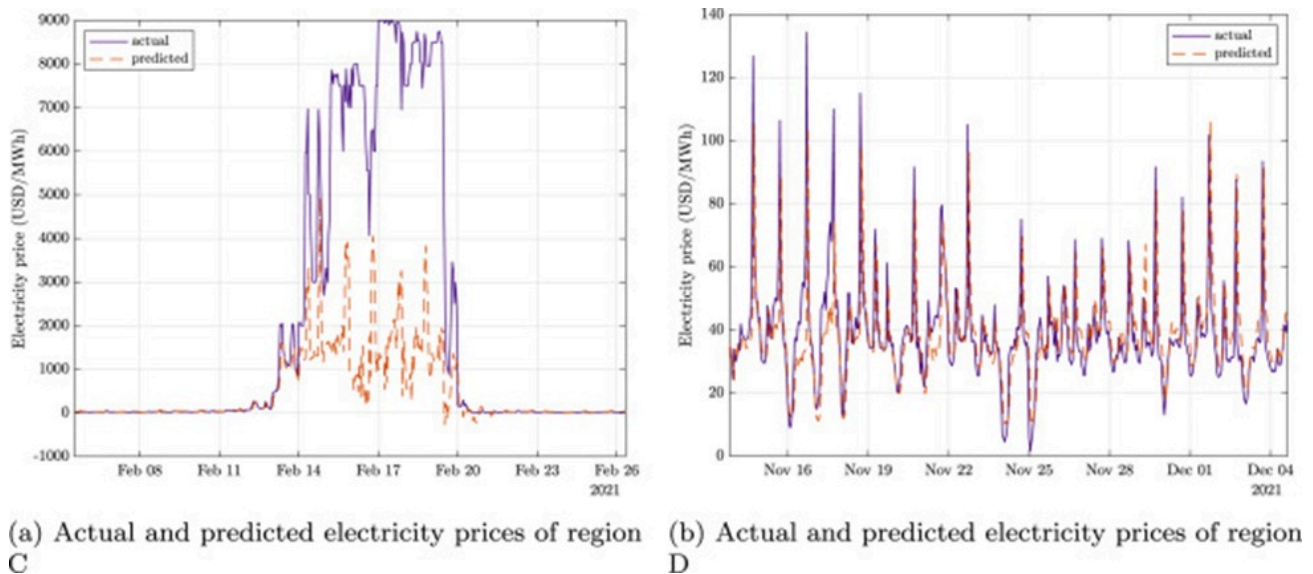


Figure 17. Actual electricity prices versus LSTM model forecasts for region C and region D.

Table 8. Performance of the proposed algorithm for the ERCOT market in region C of 2021.

Time range (from/to)		Sub-region	<i>Cgood</i>	<i>Cbad</i>	<i>Icorr</i>	<i>lincorr</i>	TN	TP	FN	FP
5 Feb, 13:00	7 Feb, 14:00	C1	93.54	5.26	60	40	1	29	2	18
7 Feb, 15:00	11 Feb, 18:00	C2	74.47	11.32	41	59	6	35	12	47
11 Feb, 19:00	15 Feb, 22:00	C3	22.22	93.41	87	13	85	2	7	6
15 Feb, 23:00	20 Feb, 02:00	C4	NA	99.00	99	1	99	0	0	1
20 Feb, 03:00	24 Feb, 06:00	C5	55.26	43.55	48	52	27	21	17	35
24 Feb, 07:00	26 Feb, 08:00	C6	78.38	7.69	60	40	1	29	8	12

Table 9. Performance of the proposed algorithm for the ERCOT market in region D of 2021.

Time range (from/to)		<i>Cgood</i>	<i>Cbad</i>	<i>Icorr</i>	<i>lincorr</i>	TN	TP	FN	FP
13 Nov, 20:00	17 Nov, 23:00	96.77	44.73	77	23	17	60	2	21
18 Nov, 00:00	22 Nov, 03:00	92.75	48.38	79	21	15	64	5	16
22 Nov, 04:00	26 Nov, 07:00	94.94	56.52	85	15	10	75	2	13

26 Nov, 08:00	30 Nov, 11:00	96.05	41.67	83	17	10	73	3	14
30 Nov, 12:00	04 Dec, 15:00	85.91	51.72	76	24	15	61	10	14

Conclusion

In practice, developers of ML-based EPF models typically emphasize the model's overall performance, yet in contexts where individual prediction errors carry significant financial consequences, stakeholders such as portfolio managers require a more granular assessment of prediction reliability. Most prior work has prioritized improving aggregate metrics, leaving a gap in understanding when specific forecasts can be trusted. To address this, the present study introduces a Trust algorithm that leverages both local and global XAI techniques to assign a trust score to each model prediction, distinguishing dependable forecasts from potentially unreliable ones. The algorithm is executed in two stages, reflecting the density and local fit principles, and is further shown to detect gradual declines in prediction accuracy over time. Its efficacy is demonstrated on CNN- and LSTM-based EPF models applied to electricity price predictions for the Italian Electricity Market and ERCOT, respectively, confirming robust performance across different models and datasets. Nonetheless, the method's reliance on predefined threshold values suggests avenues for refinement and future exploration.

List of abbreviations

AI = Artificial Intelligence

AUST = Input feature representing the past hour electricity price of Austria

CNN = Convolutional Neural Network

CNOR = Input feature representing the past hour electricity price of Italy's physical zone — Central North

COAC = Input feature representing the past hour electricity price of Corsica

CSUD = Input feature representing the past hour electricity price of Italy's physical zone — Central South

EPF = Electricity Price Forecasting

ERCOT = Electric Reliability Council of Texas

GME = Gestore dei Mercati Energetici

GREC = Input feature representing the past hour electricity price of Greece

IID = Independent and Identical Distribution

ISO = Independent System Operator

LR = Linear Regression

LSTM = Long Short-Term Memory

MAE = Mean Absolute Error

MAPE = Mean Absolute Percentage Error

ML = Machine Learning

OOD = Out Of Distribution

PFI = Permutation Feature Importance

PUN = Prezzo Unico Nazionale

SHAP = SHapley Additive exPlanations

SICI = Input feature representing the past hour electricity price of Italy's physical zone — Sicily

XAI = eXplainable Artificial Intelligence

Acknowledgements

None

Conflict of interest

None

Financial support

The work of Y. Levron was partly supported by the Israel Science Foundation, Israel, grant No. [1227/18](#). The work of K. Y. Levy was partly supported by the Israel Science Foundation, Israel, grant No. [447/20](#). The Eesti Teadusagentuur partly supported J. Belikov's work under the Estonia grant [PRG1463](#). This research was partially supported by the Israel Planning and Budgeting Committee (VATAT) for Higher Education. This research was partially supported by the Zuckerman Fund for Interdisciplinary Research in Machine Learning and Artificial Intelligence at the Technion, the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by The Nancy and Stephen Grand Technion Energy Program (GTEP), in association with the Guy Sella Memorial Project.

Ethics statement

None

Received: 28 Sep 2024

Revised: 04 Nov 2024

Accepted: 13 Dec 2024

Published online: 20 March 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and

distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Dong S, Li H, Wallin F, Avelin A, Zhang Q, Yu Z. Volatility of electricity price in Denmark and Sweden. *Energy Procedia*. 2019;158:4331-7.
<https://doi.org/10.1016/j.egypro.2019.01.788>.
- Heistrene L, Shukla Y, Kalra Y, Mishra P, Lokhande M. Impact of growing share of renewable energy sources on locational marginal prices. In: Bose M, Modi A, eds. *Proceedings of the 7th International Conference on Advances in Energy Research*. Singapore: Springer Singapore; 2021:211-21.
- Zareipour H, Canizares C, Bhattacharya K. Economic impact of electricity market price forecasting errors: A demand-side analysis. *IEEE Trans Power Syst*. 2010;25(1):254-62.
<https://doi.org/10.1109/TPWRS.2009.2030380>.
- Kath C, Ziel F. The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts. *Energy Econ*. 2018;76:411-23.
<https://doi.org/10.1016/j.eneco.2018.10.005>.
- Heydari A, Keynia F, Garcia DA, De Santoli L. Mid-term load power forecasting considering environment emission using a hybrid intelligent approach. In: *2018 5th International Symposium on Environment-Friendly Energies and Applications*. 2018:1-5.
<https://doi.org/10.1109/EFEA.2018.8617079>.
- Heydari A, Garcia DA, Keynia F, Bisegna F, Santoli LD. Hybrid intelligent strategy for multifactor influenced electrical energy consumption forecasting. *Energy Sources Part B Econ Plann Policy*. 2019;14(10-12):341-58.
<https://doi.org/10.1080/15567249.2020.1717678>.
- Mashlakov A, Kuronen T, Lensu L, Kaarna A, Honkapuro S. Assessing the performance of deep learning models for multivariate probabilistic energy forecasting. *Appl Energy*. 2021;285:116405.
<https://doi.org/10.1016/j.apenergy.2020.116405>.
- Iwabuchi K, Kato K, Watari D, Taniguchi I, Catthoor F, Shirazi E, et al. Flexible electricity price forecasting by switching mother wavelets based on wavelet transform and Long Short-Term Memory. *Energy AI*. 2022;100192.
<https://doi.org/10.1016/j.egyai.2022.100192>.
- Shah R, Shah H, Bhim S, Heistrene L, Pandya V. Short-term electricity price forecasting using ensemble machine learning technique. In: *2021 1st International Conference in Information and Computing Research*. 2021:145-50.
<https://doi.org/10.1109/iCORE54267.2021.00045>.
- Jiang P, Nie Y, Wang J, Huang X. Multivariable short-term electricity price forecasting using artificial intelligence and multi-input multi-output scheme. *Energy Econ*. 2023;117:106471.
<https://doi.org/10.1016/j.eneco.2022.106471>.
- Jedrzejewski A, Lago J, Marcjasz G, Weron R. Electricity price forecasting: The dawn of machine learning. *IEEE Power Energy Mag*. 2022;20(3):24-31.
<https://doi.org/10.1109/mpe.2022.3150809>.
- Heydari A, Majidi Nezhad M, Pirshayan E, Astiaso Garcia D, Keynia F, De Santoli L. Short-term electricity price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm. *Appl Energy*. 2020;277:115503.
<https://doi.org/10.1016/j.apenergy.2020.115503>.

Heidarpanah M, Hooshyaripor F, Fazeli M. Daily electricity price forecasting using artificial intelligence models in the Iranian electricity market. *Energy*. 2023;263:126011.
<https://doi.org/10.1016/j.energy.2022.126011>.

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.
<https://doi.org/10.1016/j.apenergy.2021.116983>.

N-Side. Time-series forecasting in energy: the need for explainable AI. 2020. <https://energy.n-side.com/blog/time-series-forecastingin-energy-explainable-ai>. Accessed September 12, 2022.

Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, Mannor S, et al. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy AI*. 2022;100169.
<https://doi.org/10.1016/j.egyai.2022.100169>.

Xu C, Liao Z, Li C, Zhou X, Xie R. Review on interpretable machine learning in smart grid. *Energies*. 2022;15(12):1-31.
<https://doi.org/10.3390/en15124427>.

Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: 31st International Conference on Neural Information Processing Systems. 2017:4768-77.

Tschora L, Pierre E, Plantevit M, Robardet C. Electricity price forecasting on the day-ahead market using machine learning. *Appl Energy*. 2022;313:118752.
<https://doi.org/10.1016/j.apenergy.2022.118752>.

Trebbien J, Gorjão LR, Praktijnjo A, Schäfer B, Witthaut D. Understanding electricity prices beyond the merit order principle using explainable AI. *arXiv*. 2022. <https://arxiv.org/abs/2212.04805>

Li W, Becker DM. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling. *Energy*. 2021;237:121543.
<https://doi.org/10.1016/j.energy.2021.121543>.

Beltrán S, Castro A, Irizar I, Naveran G, Yeregui I. Framework for collaborative intelligence in forecasting day-ahead electricity price. *Appl Energy*. 2022;306:118049.
<https://doi.org/10.1016/j.apenergy.2021.118049>.

Azam MF, Younis MS. Multi-horizon electricity load and price forecasting using an interpretable multi-head self-attention and EEMD-based framework. *IEEE Access*. 2021;9:85918-32.
<https://doi.org/10.1109/ACCESS.2021.3086039>.

Ribera TM, Lapedriza A. Can we do better explanations? A proposal of User-Centered Explainable AI. In: ACM IUI 2019 Workshops. 2019:1-7.

Li X-H, Cao CC, Shi Y, Bai W, Gao H, Qiu L, et al. A survey of data-driven and knowledge-aware eXplainable AI. *IEEE Trans Knowl Data Eng*. 2020;34(1):29-49.

Leonard JA, Kramer MA, Ungar LH. A neural network architecture that computes its own reliability. *Comput Chem Eng*. 1992;16(9):819-35.
[https://doi.org/10.1016/0098-1354\(92\)80035-8](https://doi.org/10.1016/0098-1354(92)80035-8).

Nicora G, Rios M, Abu-Hanna A, Bellazzi R. Evaluating pointwise reliability of machine learning prediction. *J Biomed Inform*. 2022;127:103996.
<https://doi.org/10.1016/j.jbi.2022.103996>.

Schulam P, Saria S. Can you trust this prediction? Auditing pointwise reliability after learning. In: International Conference on Artificial Intelligence and Statistics. 2019:1-10.

Nicora G, Bellazzi R. A reliable machine learning approach applied to single-cell classification in acute myeloid leukemia. *AMIA Annu Symp Proc*. 2021;2020:925-32.

Griffiths D. *Head First Statistics*. Sebastopol, CA: O'Reilly Media; 2009.

Rajbahadur GK, Wang S, Oliva G, Kamei Y, Hassan AE. The impact of feature importance methods on the interpretation of defect classifiers. *IEEE Trans Softw Eng*. 2021;48:2245-61.
<https://doi.org/10.1109/TSE.2021.3056941>.

di Cosmo V. Forward price, renewables and the electricity price: The case of Italy. 2015. <https://www.esri.ie/publications/forward-pricerenewables-and-the-electricity-price-the-case-of-italy>. Accessed September 6, 2022.

Energetici GM. Results of the electricity market. 2022. <https://www.mercatoelettrico.org/En/Default.aspx>. Accessed September 7, 2022.

Brown S. Fossil gas drives spiking European electricity prices. 2021. <https://ember-climate.org/insights/research/soaring-fossilgas-costs-responsible-for-eu-electricity-price-increase/>. Accessed September 6, 2022.

Daneshi H, Srivastava AK. ERCOT electricity market: Transition from zonal to nodal market operation. In: 2011 IEEE Power and Energy Society General Meeting. 2011:1-7.
<https://doi.org/10.1109/PES.2011.6039830>.

Zarnikau J, Woo C-K, Gillett C, Ho T, Zhu S. Day-ahead forward premiums in the Texas electricity market. *J Energy Mark*. 2015;8:1-20.
<https://doi.org/10.21314/JEM.2015.126>.

LCGConsulting. Energy online. 2022. <http://www.energyonline.com/Data/GenericData.aspx?DataId=4>. Accessed September 7, 2022.