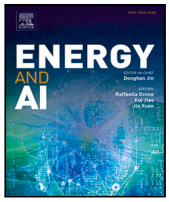
Energy and AI 9 (2022) 100169

Contents lists available at ScienceDirect

Energy and AI

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



Review

Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities

R. Machlev [a,*], L. Heistrene [c], M. Perl [a], K.Y. Levy [a,1], J. Belikov [b], S. Mannor [a], Y. Levron [a]

[a] The Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel
[b] Department of Software Science, Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonia
[c] School of Technology, Pandit Deendayal Energy University, Gandhinagar 382007, Gujarat, India

ARTICLE INFO

Keywords:
Power
Energy
Neural network
Deep-learning
Explainable artificial intelligence
XAI

ABSTRACT

Despite widespread adoption and outstanding performance, machine learning models are considered as "black boxes", since it is very difficult to understand how such models operate in practice. Therefore, in the power systems field, which requires a high level of accountability, it is hard for experts to trust and justify decisions and recommendations made by these models. Meanwhile, in the last couple of years, Explainable Artificial Intelligence (XAI) techniques have been developed to improve the explainability of machine learning models, such that their output can be better understood. In this light, it is the purpose of this paper to highlight the potential of using XAI for power system applications. We first present the common challenges of using XAI in such applications and then review and analyze the recent works on this topic, and the on-going trends in the research community. We hope that this paper will trigger fruitful discussions and encourage further research on this important emerging topic.

1. Introduction

With the evolution of deep learning (DL), better classifiers and machine learning (ML) algorithms are being developed for power system applications [1,2]. In certain scenarios, these deep learning techniques seem to have advantages compared to traditional algorithms in terms of efficiency, noise immunity, and accuracy. However, despite the evident success of such algorithms, an inherent difficulty is that since machine-learning models are often very complex, it may not be clear how or why they make certain decisions, and how they treat real-world data. Power systems planning and operation is done solely by power experts, based on their knowledge in power systems, supporting programs, and field experience, which is gathered over time. Therefore, experts in the power system field may find it hard to trust the decisions and recommendations made by machine learning based algorithms, limiting their practical use. This difficulty is especially prominent in cases that require a high level of reliability, which is common in the energy industry.

Due to this challenge, in the last couple of years new techniques and principles are being developed to improve the *explainability* of machine learning models, so that their output can be better understood. This concept is known in the literature as Explainable Artificial Intelligence (XAI) [3]. The goal of XAI is to help researchers, developers, domain experts, and users to better understand the inner operation of machine learning models, while preserving their high performance and accuracy. Various XAI techniques can be found in the literature [4,5], most of these are dedicated to DL models. Fig. 1 presents several important milestones related to both AI and power systems. It can be seen that the use of XAI in the power and energy domain has just begun.

The purpose of this work is to highlight the potential of using XAI in the context of energy and power systems. We first present the common challenges of using XAI in power and energy applications and then review and analyze the recent works on this topic and the on-going trends in the research community. We also suggest several directions for future research, that may assist in coping with the mentioned challenges, limitations, and opportunities. The specific contributions of this paper are as follows:

1. The main challenges of adopting and implementing XAI techniques in the field of energy and power systems are presented.
2. A short survey of works related to the use of XAI in power and energy applications is presented. We attempt to better understand which XAI techniques are the most common and why, and also why specific methods are used for specific applications.
3. Potential applications and future research directions related to XAI and energy systems are provided.

* Corresponding author.
E-mail addresses: ramm@campus.technion.ac.il (R. Machlev), leena80.santosh@gmail.com (L. Heistrene), michael.perl@campus.technion.ac.il (M. Perl), kfirylevy@ee.technion.ac.il (K.Y. Levy), juri.belikov@taltech.ee (J. Belikov), shie@ee.technion.ac.il (S. Mannor), yoashl@ee.technion.ac.il (Y. Levron).
[1] A Viterbi fellow.

https://doi.org/10.1016/j.egyai.2022.100169
Received 16 March 2022; Received in revised form 17 May 2022; Accepted 19 May 2022
Available online 25 May 2022

List of abbreviations

AI	Artificial Intelligence
BLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional neural network
COP	Coefficient of performance
DARPA	Defense Advanced Research Projects Agency
DeepLIFT	Deep Learning Important FeaTures
DL	Deep Learning
DNN	Deep Neural Networks
DSM	Demand side management
ELI5	Explain Like I am Five
FRLC	Fuzzy Rule Learning through Clustering
Grad CAM	Gradient class activation mapping
GRU	Gated recurrent unit
HIL	Human-In-the-Loop
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory
ML	Machine Learning
NGBoost	Natural Gradient Boosting
NILM	Non-Intrusive Load Monitoring
NN	Neural Networks
PQD	Power quality disturbance
PV	Photovoltaics
RES	Renewable Energy Sources
RL	Reinforcement Learning
RNN	Recurrent neural network
SHAP	SHapley Additive exPlanation
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

The rest of the paper is organized as follows. Section 2 provides background on XAI. Then, Section 3 presents its challenges in the energy and power domains. Later, Section 4 presents literature review of XAI for the energy and power system fields based on recent works. Section 5 discusses potential applications and future research directions, and finally, Section 6 concludes the paper.

2. Background on XAI techniques

One of the downsides of many machine learning algorithms is their "black box" nature. This means that these algorithms are extremely hard to explain, and for the most part they cannot be completely understood, even by domain experts. If users consider a model to be a black-box, they will not always trust its predictions, and therefore will be reluctant to use it. Furthermore, deep neural networks are very complex black-box models, even for AI experts, since their architecture is designed using trial and error processes, and they may consist of hundreds of layers and millions of parameters. Considering this challenge, the main goal of XAI is to allow researchers, developers, and users to better understand the results of machine learning models. One example is the DARPA XAI program, which aims to produce "glass box" models that are explainable to a "human-in-the-loop" (HIL), while preserving the model's performance [3]. Also, by adding explainability, it is believed that the generalization of ML models will be more robust allowing in the future to improve the regularization of these models. These ideas, implemented in the power systems domain, is illustrated in Fig. 2. A specific example for evaluating the performance of building energy consumption is given in Fig. 3. In this figure, XAI is used to inform the user which features are important.

In [5] the term *explanation* is defined as follows: additional information, generated by an external algorithm (or by the classifier itself), that describes the relevant features of an input instance, for a particular output. Mathematically, for a ML model f, with inputs $X \in \mathbb{R}^z$ and output prediction $\bar{y}$ for class c, an explanation map $g \in \mathbb{R}^z$ is generated to describe the feature importance, or the relevance of that input to the class output. There are various XAI techniques and approaches such as: Local Interpretable Model-agnostic Explanations (LIME) [6], SHapley Additive exPlanation (SHAP) [7], GRADient Class Activation Mapping (GRAD-CAM) [8], Deep Learning Important FeaTures (DeepLIFT) [9] and many more as presented in next surveys [5,10]. Some of these techniques, such as LIME and SHAP, are general and can be used to explain any ML model. Other XAI techniques, such as GRAD-CAM and DeepLIFT, are dedicated to DNN models in which the explanation takes the inner operations of the model, such as activation functions and weights, into consideration. As can be seen in Section 4, the most common approaches in the power and energy applications are SHAP and LIME.

There are two main scopes of explanation that XAI methods focus on. The first scope is a "local explanation", in which the input for the XAI is a single instance from the data, that is, the explanation map g is generated each time for an individual data point $x \in X$. The other scope is "global explanation", in which the goal is to understand the entire model. Mostly, this is done by using a group of data instances, and generating an explanation map g based on this group of inputs. A high-level diagram for local and global explanations is illustrated in Fig. 4. Moreover, Table 1 clarifies the difference between local and global XAI models.

Another important concept in XAI is how to integrate the explanation into the model. The explanation may be used as part of a specific ML model, or it may be applied to any model as a post-process. Specifically, there are two concepts of how to integrate XAI techniques: an intrinsic (model-specific) approach or a post-hoc (model-agnostic) approach. In an intrinsic approach, the explainability is integrated into the ML model architecture as part of the training process and cannot be transferred to other architectures. In a post-hoc approach, the XAI method is not dependent on or related to any architecture and can be applied to any trained ML model. Currently, there is significant research interest in developing model-agnostic explanations, while model-specific explanations are being used less often. A high-level diagram for intrinsic and post-hoc XAI is illustrated in Fig. 5.

To conclude the above, different scopes of XAI explanations can be useful for different purposes. Local explanations can help one understand why a specific decision was made, and can increase the user's trust in specific examples. Global explanations can help one understand complete systems and can be used to optimize them based on what the model learned. Intrinsic methods can be useful if one is training a new model, for which understanding is paramount, but post-hoc methods allow to leverage already trained models, or to use proven ML techniques. In the context of energy and power systems, there are different opportunities for using each of these techniques, as explained in Sections 4 and 5.

3. Challenges and limitations of XAI in the energy and power systems domain

There are various challenges and limitations that need to be addressed when implementing XAI for power system applications. Some of the challenges are based on [5], and extended. A summary of the challenges and limitations is presented in Table 2.

One of the main challenges is how to use models that have both high performance and transparency. Generally speaking, accurate models

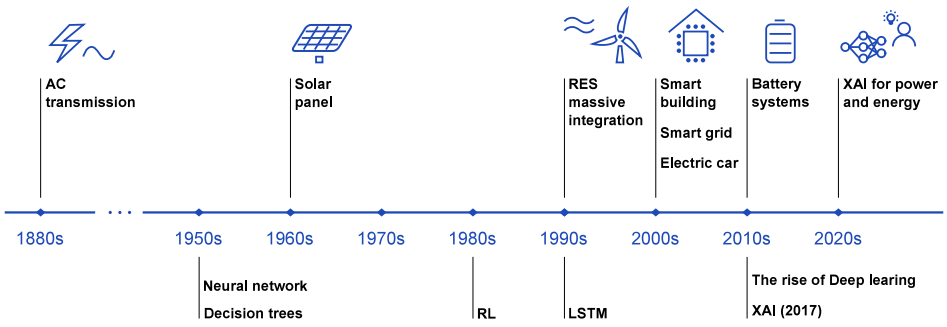


Fig. 1. The path toward XAI for power and energy systems.

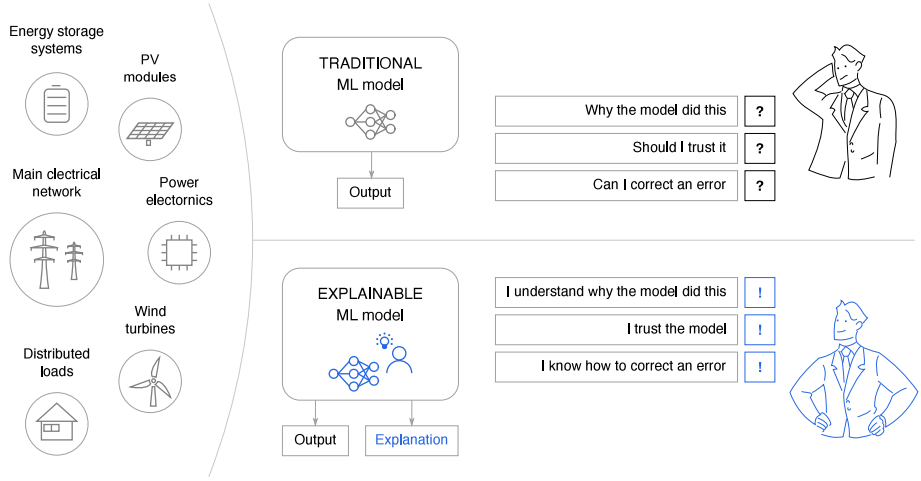


Fig. 2. Concepts of XAI for general power system applications.

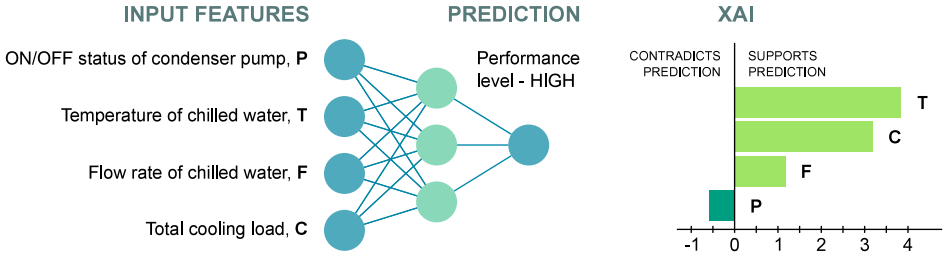


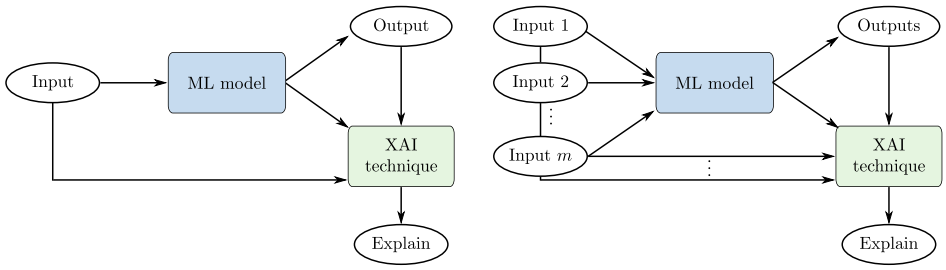
Fig. 3. Example of a classifier model with XAI for evaluating the performance of building energy consumption. The XAI technique informs the end-user which features are more relevant and important to the decision.

Table 1
Difference between local and global XAI methods.

Scope of explanation	Local	Global
Goal	Understand a single decision of the model	Understand the entire model
Input	Individual data point	Group of data instances
Explanation	Feature correlations and importance for the output of individual data point	The feature attributions of the model outputs as a whole
Familiar method	Heat-maps	Tree-based models

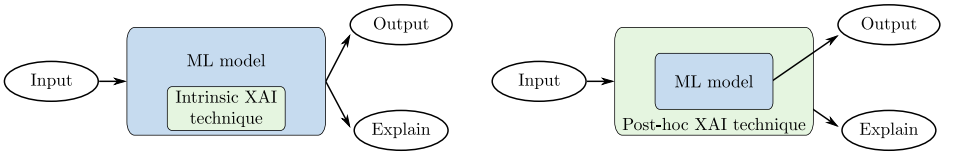
are generally more complex and hard to understand. This general trade-off is of particular importance in the power systems field, since a typical user will usually demand both high performance and an accurate explanation, in order to have a high level of trustworthiness.

In addition, a main limitation in XAI is the lack of standardization and clear definitions. Currently, while a few works and surveys define what explainability is, there is still no consensus on a specific definition for XAI and explainability. Some works focus on visualization methods, while others use the concept of feature importance or feature relevance. One of the reasons that there is no clear standard is due to the fact that there are different types of users that employ ML models and XAI, such as AI researchers, power experts, energy policy-makers, and



(a) Local explanation (b) Global explanation

Fig. 4. Conceptual illustration of locally (left) and globally (right) explainable algorithms.
Source: Adopted from [5].



(a) Intrinsic explainability algorithm (b) Post-hoc model explainability algorithm

Fig. 5. High-level illustration of intrinsic (left) and post-hoc (right) model explainable algorithms.
Source: Adopted from [5].

consumers, all of whom implement machine learning techniques for different needs, and based on varying levels of abstraction. Another important question is how to define an optimal explanation. In many applications, there is no well-defined method to provide a clear and optimal explanation. Without the ability to define this, the use of XAI might not be clear, and the outputs of the XAI algorithm may also be hard to interpret. Furthermore, while the data typically contains classification labels for the training process, usually ground-truth explanations are not provided. Without having ground-truth explanations, the XAI outputs will not have a benchmark for comparison. Accordingly, when creating a database for any energy or power application, ground-truth explanations should be given if possible. Nonetheless, as mentioned above, it is not easy to define nor to collect this information.

Another limitation of XAI techniques is the lack of evaluation metrics of the explanation quality. Even if a clear definition of explainability can be provided, it is desirable to have an evaluation metric of how "explainable" a model is. These metrics should measure an *explainability score* for each XAI technique, on any classifier. The score value is based on how much the estimated explanation from the XAI techniques is similar to the ground-truth explanation. In contrast to many other common ML tasks, there are many applications in the power and energy domain where a correct explanation can be defined unequivocally. As an example, consider the task of detecting and classifying abnormal events in a power grid. In this example, an explanation can be specified as the time at which the event appears or as the reason for the event. Thus, the explanation can be expressed as a binary vector that represents the presence or absence of the event at each sample time. The score of an evaluation metric uses this correct explanation for comparison with a XAI output. In cases where the explanation cannot be defined unequivocally, establishing objective metrics for any user will be hard to define.

An additional notable limitation is information security, which is critical in energy and power system applications. When using XAI techniques to explain a ML model's decision, the confidentiality of this model might be compromised. Therefore, any information revealed by XAI techniques may be exploited for generating effective malicious adversarial attacks to confuse the model. These attacks can manipulate the model by feeding the system specific information, which leads to a different output. Based on the XAI outputs, these attacks can be very effective since they can find the minimum changes that should be applied to the input data in order to change the decision of the ML model. The consequences of such attacks on power grids or energy management systems could be catastrophic.

Another challenge is that current techniques provide explanations that are designed for AI experts instead of power systems experts. Currently, the most advanced XAI algorithms were developed by computer scientists and AI researchers. Therefore, the common way to provide an explanation is by using heat-maps, which provide good intuition but do not always contain enough information for the user. In many applications, it is possible that more sophisticated presentation of the XAI outputs may lead to a more useful explanation. Accordingly, it may be beneficial to form collaborations between power and energy domain experts, cognitive scientists, and others to create efficient and dedicated XAI techniques for these applications. Furthermore, these techniques should be optimally suited to the relevant user. For example, techniques that explain recommendation models for power grid operation should benefit from embedding prior knowledge of the system operators.

A critical challenge is how to prevent XAI methods from outputting misleading explanations and how to provide trustworthy recommendations. Explanations can increase the user's trust, but in the long term, the outputs of the models may not make accurate recommendations. Such a risk might cause users to develop incorrect confidence and trust mistaken results. Furthermore, believing in wrong explanations or predictions may encourage users to invest confidence in ineffective or unsafe models, which later can lead to disasters. An example can be a model agnostic approach, which can be used for any classifier and whose outputs might be too general and not dedicated to the specific application, thus their reliability should be doubted. Accordingly, it might be better to use an intrinsic model that is designed for a specific application. Another question is whether one should trust a model based on good explanations of available data instances, or should one trust a model only when it is verified on the entire dataset. Gaining trust from a locally explainable algorithm requires verification of many different inputs, while the approach of a globally explainable algorithm can reach reasonable explanations based on the entire dataset. These comparisons of intrinsic versus post-hoc and local versus global approaches for power and energy applications should be further studied and examined.

Table 2
Open challenges and limitations of XAI methods for the energy and power systems domain.

Category	Challenges and limitations
Tradeoffs	✓ Both the explanation and the performance must be accurate
Standardization	✓ Lack of agreement on definition for XAI and explainability ✓ Different types of users–AI researchers, power experts, energy policymakers, and consumers ✓ Ground-truth explanations should be given
Evaluation metrics	✓ Objective metrics should be used if explanation cannot be defined ✓ Explainability score need to be measured and evaluated
Security	✓ The confidentiality of ML models might be compromised ✓ Malicious adversarial attacks might be generated using information revealed by XAI techniques
Users	✓ Collaborations between energy experts and other domains' experts ✓ XAI techniques should be optimally suited to different types of users
Recommendations	✓ Prevent XAI methods from outputting misleading explanations ✓ Users might develop incorrect confidence and trust mistaken results

4. Survey of XAI for energy and power system applications

This section reviews and analyzes recent works and trends related to XAI in power and energy applications. In Section 4.1 the latest works of XAI for critical power grid practices are presented. Then, in Section 4.2 papers that examine the use of XAI for forecasting renewable energy generation, energy demand, and other applications are reviewed and analyzed. Later, Section 4.3 focuses on XAI for building energy management applications. Finally, Section 4.4 presents a summary of the overall literature survey with an attempt to better understand the unique characteristics, trends and common uses of XAI within the power and energy research community.

4.1. Explainable AI models focusing on power grid applications

Due to the growing penetration of renewable sources liberalization of electricity markets, and increasing presence of direct current transmission systems, the power grid's vulnerability is increasing. Recently, advanced AI techniques have shown great promise in dealing with stability analysis, and other control related aspects, but the practical implementation of these AI techniques has been hindered by the lack of transparency of these black-box models [11].

Some works that focus on explainability of grid security assessment models were published in the last couple of years. Work [12] uses decision trees for classifying system operating conditions as stable or unstable. The decision tree represents the grid security rules while its tree depth depicts interpretability. This work finds a trade-off between the model accuracy and interpretability through their modified optimal classification tree. To the best our knowledge, this is the only work that uses an intrinsic approach for explainability in the power system domain. Other works discussed below use model agnostic approaches for explainability. For instance, in [13] a Gated Recurrent Unit (GRU) is used for classifying grid security conditions. The general framework of this transient stability assessment model includes offline training with GRU and an interpretable decision tree model. This tree acts as a surrogate model for explainability, and can be implemented online for enabling preventive control measures. While a visualization approach is used in [13], a knowledge extraction and influence approach is applied in [14] for explainability. Here, transient stability status is modeled using Extreme Gradient Boosting (XGBoost), and its performance is evaluated through two evaluation metrics—missing alarm rate and false alarm rate. Similarly, transient stability assessment is also discussed in [15]. Here, transient stability is modeled using a deep belief network, and is explained through a local linear interpreter model. Both fidelity and interpretability are discussed in this work. Fidelity ensures that even though the model may not have good global performance, it remains faithful for specific fault scenarios at the local level. Finally, [16] uses Shapley values with heat maps for explaining the ML model used for transient stability assessment.

A similar instability event in power grids is frequency deviations. Researchers have often used ML models to predict the dynamics of frequency deviation phenomena. The authors in [17] have used a gradient boosting tree technique to predict frequency stability indicators. To make the predictions made by this ML model more understandable to the operator, they employed the SHAP technique. In later works from the same authors, SHAP provides insights into the role of driving factors such as forecasting errors and ramping rate of generators on frequency deviations for different grids [18]. In another work, the same authors have also analyzed and explained secondary control activation with SHAP. They used automatic load frequency controllers to provide the necessary reserve that would restore generation load imbalance [19].

Voltage instability is yet another issue in power grids, and grid operators often resort to load shedding to prevent excessive drop in voltage. Authors of [20,21] have applied SHAP explainability to a Deep Reinforcement Learning model for the implementation of proportionate load shedding in undervoltage conditions. The deep-SHAP method is used to increase the computational efficiency of the XAI model in [21]. Its input layers consist of a backpropagation strategy that calculates the significance of all the input features and a softmax layer with a probabilistic representation method. The output of this model explains the model predictions through a visualization layer and feature importance layer. Both global and local explanations are discussed in this work. A summary of XAI techniques used for system stability applications is given in Table 3.

Recently, some works discuss identification of different fault scenarios in power grids. For example, in [22] grid disturbance events such as generation tripping, line tripping, system oscillation status, islanding, and load shedding events are identified through a Long-Short Term Memory (LSTM) model. The deep SHAP method, which involves a combination of DeepLIFT and SHAP values, explains the LSTM model. Insights into misclassification using explainability are also discussed in this work, which is an important concern when implementing an AI model for any application. A frequency event which is taken from work [22] is presented in Fig. 6. In this figure, an explanation based on SHAP values is presented for classification of generator tripping. The magenta time-steps are the features that contribute the classifier decision and the black time-steps are the features that does not contribute to the decision. Accordingly, the parts of the time-series which have a greater contribution to the classifier decision are time-steps 4 and 5. These time-steps are the frequency's downfall in the middle of the time-series. Meanwhile, time-steps 7 and 8 do not have a relevant impact on the decision.

In [23] power quality disturbance (PQD) classification is discussed. Here, performance of various XAI techniques are compared for individual disturbances at the validation stage, and the best among them is selected to explain the inference model. To the best of our knowledge, this is the only work that provides a definition for the correct explanation of PQD classifiers, referred to as a 'ground truth' by the authors of this work. The same work also proposes a PQD explainability score which makes it possible to evaluate the explanation of each XAI technique and DL classifier. To demonstrate the above, Fig. 7, which is taken from [23], presents the explanation of CNN classifiers output for two XAI techniques when the input is Sag signals. For both pictures, the XAI output is a heat-map in which the colors are aligned to the relevance of the features in the decision. In Fig. 7(a) a sag signal is

Table 3
Summary of XAI techniques used in various system stability applications.

Ref.	Power system application	AI model	XAI approach	Scope
[12]	Security assessment	Decision tree		Intrinsic & Global
[13]	Security assessment	Gated Recurrent Unit	Tree regularization	Global
[14]	Post fault transient stability assessment	Extreme Gradient Boosting	Decision rules & LIME	Global & Local
[15]	Transient stability assessment	Deep Belief Network	Local linear interpreter	Local
[16]	Transient stability assessment	Machine learning	Shapely values	Local
[17]	Predicting frequency stability indicators	Gradient Tree Boosting	SHAP	Local
[18]	Predicting deterministic frequency deviations	Gradient Tree Boosting	SHAP	Global & Local
[19]	Activation of secondary control power	Gradient Tree Boosting	Partial dependency plots & SHAP	Local
[20]	Undervoltage load shedding	Deep Reinforcement Learning	SHAP	Global & Local
[21]	Undervoltage load shedding	Deep Reinforcement Learning	Deep SHAP	Global & Local

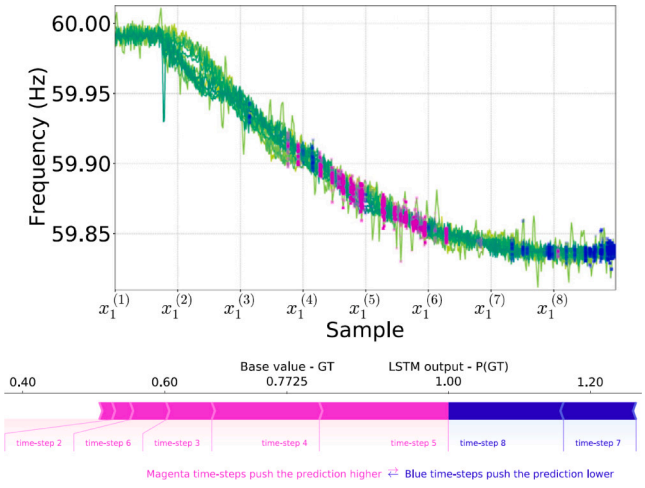


Fig. 6. Frequency signal detected as generator tripping with feature contribution highlighted.
Source: Taken from [22].

explained by occlusion-sensitivity in which the explanation is related, almost perfectly, to the entire disturbance. In Fig. 7(b) a real-life sag signal is explained by GRAD-CAM in which some of the strongest features are not related directly to the disturbance.

Another example is the case of diagnosing faults observed in a power transformer, as shown in [24]. Timely detection of such incipient faults can help avoid deteriorating impact on individual grid components and the overall grid as a whole. In this work, dissolved gas analysis is done using a neural network based classification model. This model utilizes the correlation of the dissolved gases with the fault type. Then, SHAP is used to select the most compelling features best suited for the proposed classification model. Another issue is faults that relate to photovoltaic (PV) installations. An explainable model is proposed in [25,26] to detect and diagnose incipient faults in PV installations, which would help plan routine maintenance operations, and help in troubleshooting. A hybrid modeling approach is adopted in [25] consisting of both model-based and data-driven methods for fault detection. The LIME approach used in this work helps the user understand why the XGBoost model predicted the existence of a fault. With the practical implementation of the explainable fault detection system on an edge node, an advanced version of the same application is discussed in [26]. An edge-based fault detection system increases scalability and reduces data exchange and data security concerns. The edge node services described in the model proposed in [26] provide fault explanations with time stamps. A summary of the results discussed above, detailing XAI techniques used in grid event detection applications, is given in Table 4.

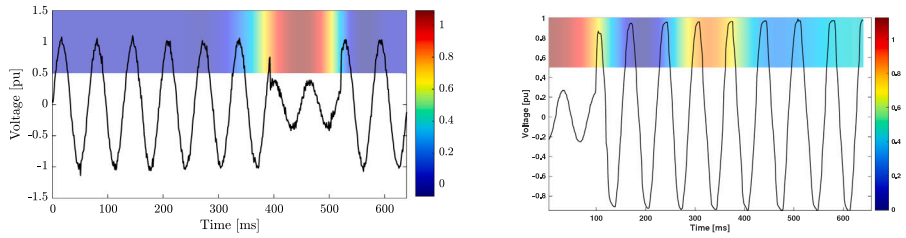
Table 4
Summary of XAI techniques used in various grid event detection applications.

Ref.	Power system application	AI model	XAI approach	Scope
[22]	Event identification	LSTM	Deep SHAP	Global & local
[23]	Power quality disturbances classifier	CNN	Occlusion sensitivity, GRAD-CAM & LIME	Local
[24]	Fault diagnosis of power transformers	XGBoost	SHAP	Local
[25]	Detection & diagnosis of incipient faults in PV panels	XGBoost	LIME	Local
[26]	Edge node based explainable incipient fault detection system for PV	XGBoost	LIME	Local

4.2. Explainable AI models for the energy sector

ML models are used in many applications related to the energy sector. These applications have a significant impact on planning and operation decisions, and therefore it is essential to have explainable XAI models that can gain the trust and confidence of the end-user [27].

Lately, many works have focused on using XAI for renewable energy forecasting. As an example, an XGBoost model is used for predicting PV power generation in [28]. The authors have used ELI5 (a Python library which provides feature importance for machine learning models) to explain this XGBoost model. Based on the explanation of ELI5, the unimportant features are removed from the input dataset. However, the decrease in the input dataset size came at the cost of reduced model accuracy. This idea is extended in [29] by using three XAI techniques to analyze the feature importance more elaborately from different perspectives. LIME, SHAP, and ELI5 are used to explain the PV power predictions made by the random forest regressor model.



(a) occlusion-sensitivity (b) GRAD-CAM

Fig. 7. XAI for PQD classification. Explanation of two different Sag signals using (a) occlusion-sensitivity and (b) GRAD-CAM on DL classifiers. The classifiers predict the signal as 'Sag'.
Source: Taken from [23].

Based on the feature importance analysis from different XAI techniques, the authors concluded that excluding two features from input data is reasonable without compromising on the model accuracy. This exclusion was validated by the forecasting results, wherein it was observed that the removal of unimportant features improved the model accuracy compared to the baseline model. The purpose of XAI in both these works was to improve ML model design for a specific application.

Residential consumers with PV sources need accurate PV power prediction models to formulate an intelligent demand response strategy. This issue is addressed through a tree-based interpretable machine learning model in work [30], wherein feature attribution methods are also used to discover features with higher impact. Two different neural networks (NN) models, a binary classifier NN model and a regressor NN model, have been used for a similar prosumer application in [31]. The binary classifier predicts whether solar power generated will be greater than the load or not. The regressor model gives a continuous value prediction for PV power output and load demand. Either of these predictions helps the prosumer to plan curtailment or import as needed. An attribution method with integrated gradient, expected gradient, and DeepLIFT approaches is executed to interpret both—the classifier NN model and the regressor NN model. Additionally a Bayesian NN captures the uncertainty associated with failure in prediction. Explainability applied to PV generation forecast is also seen in [32], wherein the explainability is applied to the prediction interval of a probabilistic forecast. In this work, the initial probabilistic forecast of Natural Gradient Boosting (NGBoost), Gaussian process, and Lower Upper Bound Estimation algorithm is compared. NGBoost is found to be better than the other two methods in terms of probabilistic forecasting performance and average training time. So, in [32], probabilistic forecasts from NGBoost form the first stage of the proposed model, while the SHAP explanations make up the second stage. Global feature importance plots based on SHAP values, interaction plots, and force plots are used to explain the predictions of the PV power forecasting model.

Renewable energy applications discussed until now were based on predicting PV power. The following works use XAI techniques to explain ML models that predict solar irradiation. Work [33] discusses explainability through an intrinsic model for solar irradiance forecasting. The authors use a direct explainable neural network—a feedforward network wherein the output can be mapped with the input through a non-linear model. In study [34] the Fuzzy Rule Learning through Clustering (FRLC) model is proposed to predict solar irradiation, and this FRLC model is interpreted through a knowledge-based approach. Membership function plots for different features and similar linguistic methods are implemented to interpret the FRLC model predictions. In work [35], permutation feature importance, SHAP, and feature dependency analysis are used for explaining the predictions of a solar irradiance forecast model.

The next work focuses on XAI for general applications (other than renewable energy forecasting) in the energy domain. Work [36] uses XAI techniques for interpreting load forecasting models. The authors use a visualization approach and an influence approach to find the best-suited subset of features that could improve model accuracy. Electricity generation mix is predicted using the Light Gradient Boosting model in [37]. SHAP is used on this model to understand the relevance of different features, and accordingly, the appropriate input dataset is provided to the AI model to improve model accuracy. In study [38] a Bidirectional LSTM (BLSTM) is used on the time series of load, wind, net import/export, etc. An attention mechanism explains how the temporal characteristics can be used to improve the accuracy of this BLSTM model.

Furthermore, AI has been extensively used in the power electronics domain for applications such as prediction of remaining useful life of supercapacitors and abnormality detection of inverters [39]. Explainability has been applied to one such intelligent controller used for inverters in [40]. The authors present a data-driven explainable AI controller for a grid-connected voltage source inverter. Using a conditional probability density function, the impact of individual features on the output is assessed, and erroneous data is identified through interpretations obtained from conditional entropy, conditional entropy weights, and outlier determination. Removal of these erroneous data from the training set improves the quality of the model.

All the above works are further summarized in Table 5.

4.3. Explainable AI models for energy management in buildings

This subsection discusses the explainability of AI models that are used to forecast the energy consumption of buildings, or are generally related to energy efficiency in buildings.

In [41] an XAI technique is applied to a long-term prediction model for forecasting the cooling energy consumption of buildings. Here, SHAP provides insights that black box models cannot offer. Another explainable long-term prediction model was introduced in [42], wherein the predictions of annual building energy performance are studied using explainable models. This work explains the QLattice algorithm, a ML model used for forecasting building energy performance. Building energy modeling in terms of energy usage is discussed in [43]. This model is explained using a visualization approach. Localized decision tree models are used as surrogate models to explain decisions about a specific region in the explanation space. Additionally, XAI techniques have been applied to short term forecasting models that predict the buildings' energy performance. The coefficient of performance (COP) of a building is based on the building's total cooling load. In [44], the high/low status of COP is predicted through k-means clustering. XAI techniques are applied to several different machine models in this work. It is observed that while specific models stood out for their accuracy, they did not perform well under trust metrics. Research work on a similar theme is seen in [45] wherein the DL models, used for predicting the performance of irregular dew point coolers, are explained using SHAP values.

Another approach, using attention mechanisms for explainability, is seen in [46,47], and [48]. In [46] building energy demand is predicted through a type of RNN model called as sequence-to-sequence model. This model is explained using feature importance based on weights

Table 5
Summary of XAI techniques used in energy applications.

Ref.	Energy application	AI model	XAI approach	Scope
[28]	PV power generation forecasting	XGBoost	ELI5	Global
[29]	PV power generation forecasting	Random Forest Regressor	SHAP, LIME, ELI5	Global & Local
[30]	PV power generation forecasting to plan demand response strategy for prosumers	Tree-structured self organizing map & XGBoost	Feature attribution from weight, gain and cover & SHAP	Global & local
[31]	Predicting whether solar power produced will suffice the load demand of the prosumer	Binary classifier NN & regressor NN	SHAP, Integrated and Expected Gradients & DeepLIFT	Global & Local
[32]	Probabilistic PV power generation forecasting	NGBoost	SHAP	Global & Local
[33]	Solar irradiance forecasting	Direct explainable Neural Network		Intrinsic, Global
[34]	Solar radiation forecasting	Fuzzy Rule Learning through Clustering	Knowledge-based approach using linguistic methods	–
[35]	Solar radiation forecasting	Light Gradient Boosting	Permutation Feature Importance and SHAP	Global & Local
[36]	Load forecasting	XGBoost	SHAP & partial dependence plots	Local
[37]	Predicting electricity-generation mix	LightGBM	SHAP	Local
[38]	Predicting area control error in renewable rich grids	Bidirectional Long Short Term Memory	Attention Mechanism	Global
[40]	Controller for grid connected VSI	–	Conditional Entropy	Global

of input features and through an attention mechanism. Interpretability through the attention mechanism is discussed in depth in [47]. The deep learning technique used in this work is an encoder–decoder model with LSTM models in sequence. When this model is combined with a self-attention mechanism, it ensures that accuracy and explainability form an optimal trade-off. In [47], attention based on hidden layers and attention based on input features have been explored. Similarly, in [48] a building's cooling load demand is predicted, with a 24-hour ahead prediction horizon, using a RNN structure with an attention mechanism for interpretability. In this work, the soft attention mechanism is used for two case studies—the univariate and multivariate models. It is observed that with different RNN structures (LSTM and GRU), the training time per epoch increases, but the error reduces. The attention matrix formed from the attention vector explains the contribution of each input feature. The density of the attention matrix can be used to get insights into the model and its dependencies.

Furthermore, benchmarking of buildings using explainable AI is discussed in the works [49,50]. Clustering similar buildings based on usage patterns is the prerequisite for benchmarking the performance of buildings. In [49], buildings are classified based on usage pattern and then building performance is compared. This work focuses on the explainability of the clustering models trained on smart meter data obtained from non-residential buildings. In [50], XGBoost is used for benchmark analysis, and the interaction among different input variables and their impact on the score is explained using SHAP.

It is essential to make the users of the building an integral part of the Building Energy Management System. This can be achieved by encouraging them to change their usage patterns such that the building performance can be improved. Authors of [51] have considered a smart building infrastructure wherein explainable AI models assist the building users to make proper decisions in response to building performance indicators. Change in the users' consumption can be achieved if the users find a reason or incentive to alter their behavior. Authors of [52] have proposed the segmentation of different users through clustering to decide the proper motivation for each group of users. This clustering model is explained using the Graphical Lasso approach to understand feature correlation. A similar instance of applying explainability for clustering models is seen in work [53] for identifying rules of low energy solutions for different cities. A summary of all the above XAI techniques discussed in this subsection is given in Table 6.

4.4. Survey summary

The literature survey presented in the above subsections clearly shows that engineers and researchers are using XAI techniques more and more in energy and power system applications. Since 2019, we see a steady growth in the number of publications describing XAI techniques for these applications, as shown in Fig. 8. The keywords we used for searching for papers are summarized in Table 7. We have classified the research papers surveyed in our work based on their application domain. A diagrammatic representation of the same is shown in Fig. 9(a). The majority of publications have been from the building energy management section, and for applications that tackle grid vulnerability issues. The works have also been categorized based on the ML model used. It is found that most XAI models have been applied on Gradient Boosting (GB) techniques, especially XGBoost which has been most popular GB algorithm among researchers. Only 28% of the XAI techniques surveyed were used with deep neural networks models, an indication that DL with XAI for power and energy applications is still not widely used. This information can be seen in Fig. 9(b).

Fig. 10 shows a classification of papers based on the characteristics of the XAI model. As mentioned in Section 2, and as observed in Fig. 10, very few research works have implemented ML models with intrinsic explanation capabilities. In contrast, most works use the post-hoc model agnostic approach. According to [54], this approach can be further classified into: (i) visualization approach, (ii) knowledge extraction approach, (iii) influence methods, and (iv) example-based explanation. To the best of our knowledge, no research related to the power and energy sector has adopted the last category. It is also observed that

Table 6
Summary of XAI techniques used for building energy management applications.

Ref.	Building energy applications	AI model	XAI approach	Scope
[41]	Predicting long term cooling energy consumption in buildings	XGBoost	SHAP	Local
[42]	Predicting long term building energy performance	QLattice	Permutation feature importance	Local
[43]	Predicting energy usage models	–	Partial dependence plot, Surrogate model	Local
[44]	Predicting coefficient of performance of the cooling system	SVM, MLP, XGBoost, RF	LIME	Local
[45]	Performance forecast of irregular dew point cooler	Deep Neural Network	SHAP	Local
[46]	Short term forecasts of building energy consumption	Sequence to Sequence model	Attention mechanism & Feature importance	Local
[47]	Short term forecasts of building energy consumption	Encoder Decoder model with LSTM sequence	Attention mechanism	Local
[48]	Short term forecasts of building energy consumption	Encoder Decoder model with RNN sequence	Attention mechanism	Local
[49]	Identifying usage pattern and building energy performance	Classifier using ML	Correlation among temporal features	Local
[50]	Benchmarking building energy performance levels	XGBoost	SHAP	Local
[51]	Improving HILs' energy usage pattern through smart building infrastructure	LSTM	Feature correlation using graphical Lasso & Granger causality	–
[52]	Clustering for deciding incentive for improving energy efficient response	K-means algorithm (unsupervised clustering)	Feature correlation using graphical Lasso & Granger causality	Local
[53]	Identifying design strategies for low energy building solutions	Axis parallel Hyper rectangle algorithm	Rule based approach	Global

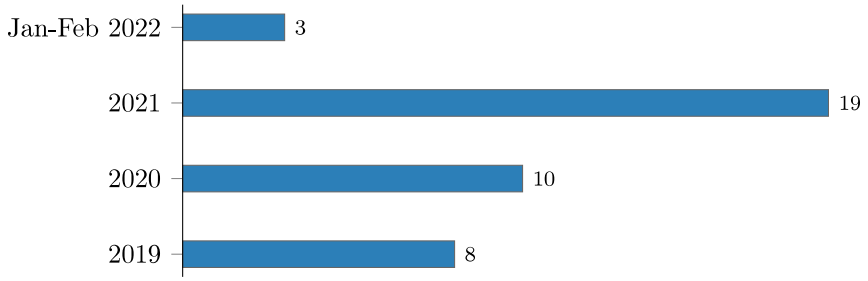


Fig. 8. Year-wise distribution of publications on XAI techniques in energy and power system applications.

the majority of research papers have implemented various influence methods, such as feature importance techniques, layer-wise relevance propagation, and sensitivity analysis. Further distribution of these influence methods can be seen in Fig. 10. These statistics indicate the vast scope of research opportunities for XAI that are yet unexplored in the power and energy domain.

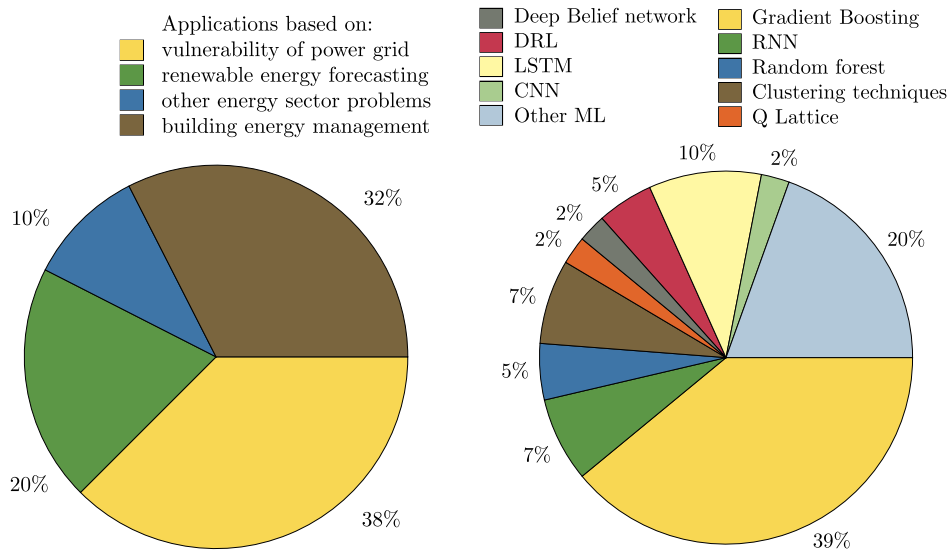
5. Opportunities of XAI for energy and power system applications

In this section, some opportunities of XAI for energy and power systems are identified by looking at applications that use ML but have not yet considered XAI. Such applications can benefit significantly from the models' decisions becoming transparent.

The suggested research directions outlined in the section are summarized in Table 8. These are only some suggestions which are used to demonstrate how there are many possible avenues for using XAI in power systems and energy research.

5.1. Optimal energy management and control

Modern electric grids are increasingly decentralized and deregulated. Therefore, it is often impractical to study grids from the perspective of a single entity with unlimited information and control span. Furthermore, the statistical behavior of loads and renewable energy sources are becoming more complex, causing traditional control methods to be less effective. The result of these issues has driven research in control using ML techniques for problems of energy management



(a) Based on power and energy applications

(b) Based on machine learning algorithms

Fig. 9. Classification of papers based on the type of application and the machine learning model.

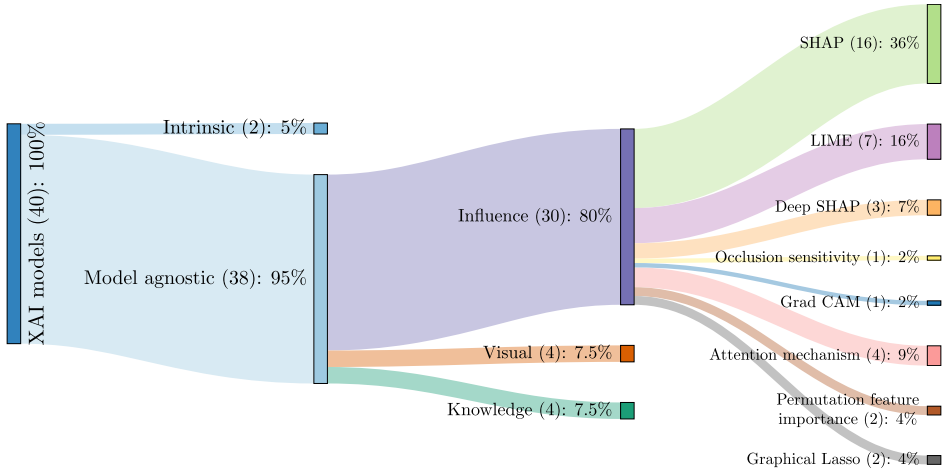


Fig. 10. Classification of papers based on the characteristics of the XAI model used.

Table 7
Keywords for different application areas.

Section		Search keywords
4.1	Vulnerability of power grid	'explainablility'/'interpretability' AND 'transient stability'/'frequency'/'voltage'/'power quality' AND 'fault diagnosis' AND 'SHAP'/'LIME'/'feature importance' AND 'XGBoost'/'deep learning'
4.2	Energy sector applications	'explainablility'/'interpretability' AND 'forecasting' AND 'renewable energy'/'generation AND 'energy'/'load demand' AND 'SHAP'/'LIME'/'feature importance' AND 'regressor'/'temporal'/'deep learning'
4.3	Building energy management applications	'explainablility'/'interpretability' AND 'forecasting' AND 'building energy' AND 'SHAP'/'LIME'/'feature importance' AND 'regressor'/'temporal'/'deep learning'

such as work [55]. However these techniques have the disadvantage of being uninterpretable, which may make power system engineers wary of relying on them. Applying XAI can increase the reliability of these ML models, and may lead to a variety of future research avenues. Next, some familiar applications in this area are presented which could benefit from the addition of explanations provided by XAI.

One such application is energy storage control, which becomes necessary with the increasing integration of intermittent renewable energy sources. These days, reinforcement learning (RL) is a field being explored for control of energy storage units [56]. At the same time, the reliability of the storage systems is essential, so it is important that any control algorithm used is trustworthy. One reason a power engineer may not want to rely on RL techniques is that the learned policy may make decisions that do not have a clear reason. Therefore, a future research direction can be to use explainable reinforcement learning for the design of controllers that power engineers can verify and understand.

Other common control applications that can benefit from using XAI are frequency control and voltage control. Recently, reinforcement learning has been increasingly employed to solve such problems [57, 58]. One reason for this trend is the increasing integration of renewable energy sources, and the introduction of "prosumers" [59]. For example, power engineers may worry that a learned policy may not cover some extreme scenarios–such as load shedding, leading to unintended and damaging behavior. Accordingly, the use of XAI can help one understand the control laws learned by the RL algorithms, and help grid

Table 8
AI applications in power systems and energy with the benefits of XAI.

Application	Gap from lack of explainability	Benefit of XAI
Energy storage control using reinforcement learning	The control laws learned may not make sense to a power system expert and may be seen as unreliable	Explaining the reasoning behind the learned control laws and providing understanding to the power expert
Frequency/Voltage control using reinforcement learning	The control laws learned may be seen as unreliable in extreme scenarios such as load shedding	Explaining control decisions in extreme scenarios which can be verified by a power expert
Energy scheduling using RL	Unclear reasoning behind scheduling decisions	Explanations of why scheduling decision was made
NILM	Lack of consumer trust in uninterpretable ML algorithms	Increase user trust by explaining reasoning behind decisions
Demand side management	Consumer worry of relying on an uninterpretable algorithm for a critical purpose, such as charging a car	Provide user with understanding and reassurance
Line fault detection	Lack of understanding of which measurements drove the decision	Show which measurements are most important to the decision, allowing better system planning
Cyber-security monitoring	Uninterpretable decisions in a critical application	Show why a measurement is flagged and allow investigation into the reasons it is considered a cyber-attack

operators employ smarter control strategies for distributed generation, and prevent frequency and voltage stability issues. Furthermore, the use of these powerful tools in conjunction with the understanding afforded by XAI methods can be a pathway for faster adoption of intelligent systems which optimally manage multiple distributed sources.

Reinforcement learning is also employed for energy scheduling applications. For example, RL is used in [60] for dynamic pricing. These algorithms do not provide reasoning of how the price is reached. If an ML model has sudden jumps in price, it is important for the system manager to know why. Nevertheless, it is vital to explain how the price of electricity is determined for the consumers, making this a good candidate field for XAI. Another example is [61] which uses RL for solving economic unit commitment. These algorithms do not explain why a unit commitment decision was made. Therefore there may be cases where the algorithm makes a decision which is unclear to the system manager. XAI can be particularly useful in this case since it allows ML methods to be more understandable, and thus provides an information source that can be used for long-term planning.

5.2. Energy consumer applications

One recent advancement in smart grid technology is consumer utilization of smart meters. Accordingly, in the last few years many works have implemented ML techniques for energy consumer applications which use information obtained from these meters such as work [62]. This information may encourage consumer energy saving behavior, improve fault detection, improve demand forecasting, enhance energy incentives, and more [63]. However, the practical usefulness of using ML for these tasks may be limited, because users may find it hard to understand the decisions made by such algorithms. Therefore, XAI can improve the user's trust for such applications. The following text show some cases to highlight the above claims:

One application of smart meters is efficient estimation of the power consumption of individual devices through load disaggregation, which is also known as Non-Intrusive Load Monitoring (NILM). In the last few years many works have implemented deep learning techniques for NILM [64]. Another application for consumer energy saving is demand side management (DSM). Deep learning has been effective in solving particular applications of this problem, such as charging electric vehicles [65]. However, an obstacle for real-world adoption is the lack of consumer trust For example, a consumer may not want to use an ML model he does not understand for charging an electric car due to a fear of the car not being fully charged when needed. Therefore, for these applications, using XAI techniques can help increase understanding and usefulness of such models to the consumer-user by providing trustworthy and simple feedback. Note that for such applications, works [66,67] provide concepts of how to make classifiers interpretable specifically for AI specialists, and not only for end-users trust improvement.

5.3. Power system monitoring

Another common application which uses ML is power system monitoring. This includes fault detection and location, identification of imbalances in power networks, and identification of cyber-security attacks. Increasing network complexity makes these tasks difficult, and therefore, machine learning methods, which are excellent at solving these types of complex pattern recognition problems, are frequently employed to do so [68]. However, because of their black-box nature, machine learning models are prevented from being used in practice for these critical applications. The use of XAI can allow to leverage these powerful models without comprising the trust needed for these critical tasks.

One possible application is line fault detection, which uses bus measurements as input features [69]. While there has been excellent results in research when using ML for this task the reasons for the predictions are not clear. Power engineers may be wary of using a model which may make decisions based on irrelevant information. The use of XAI can help power system engineers know if the prediction is based on relevant information. For example, XAI may be used to understand which measurements were used to locate the fault, or global XAI methods can be used to help system planners optimally place measurement equipment on the buses which the model considers the most important.

Another application of power system monitoring is cyber-security monitoring. ML methods are being used to detect cyber-attacks such as false data injection [70] and intrusions [71]. However, these methods do not explain why the measurements are considered a cyber-attack. Thus, operators may not be able to understand the attack and it may be hard to solve. Also, as discussed previously, understanding of these models needs to be very high due to the disastrous consequences of misclassification. XAI techniques can help with trusting these models and allow power experts to know why an event is considered a cyber-attack.

6. Conclusion

Advanced machine-learning models have recently demonstrated outstanding performance when applied to energy and power system applications. Nevertheless, power experts and users may find it hard to trust the results of such algorithms if they do not fully understand the reasons for a certain algorithm's output, and how it operates in practice. Accordingly, the goal of Explainable Artificial Intelligence (XAI) is to transform ML models so that they are more explainable and trustable. As part of this trend, XAI has been implemented in several works in the energy and power domains over the last several years. The applications we focus on in this paper have been selected following an in-depth content analysis of various sources, as detailed in Section 4.4. This analysis reveals interesting trends in the current research, and may help one understand under which conditions the different XAI techniques are used. Specifically, the data shows that SHAP and LIME are the most widely used XAI techniques. Furthermore, most of the ML models that use XAI are traditional ML algorithms, while DL models are rarely used with XAI.

Another important aspect covered in this work is the challenges and limitations of adopting and implementing XAI techniques in the field of energy and power systems. While XAI can facilitate the use of ML techniques in practice, there are still obstacles that should be considered, such as standardization, security, and incorrect confidence. In addition, potential applications and future research directions related to XAI and energy were provided. Some of these are: optimal energy management and control, energy consumer applications, and power system monitoring. One suggested future research with high potential is the use of XAI for understanding control laws learned by reinforcement learning algorithms, which can lead to faster integration of renewable energy sources and energy storage units in various applications. Moreover, another recommended direction is to implement XAI techniques with DL models since, regardless of their high accuracy, DL models are not transparent and thus are hard to trust without explanations.

To conclude, this work presents many examples and opportunities of how XAI can be useful in the energy and power systems domain. We believe that despite the challenges mentioned above, XAI techniques have significant potential to explain the decisions of machine learning algorithms, which are increasingly being used nowadays in the energy and power systems community.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of Y. Levron was partly supported by Israel Science Foundation, grant No. 1227/18. The work of K. Y. Levy was partly supported by Israel Science Foundation, grant No. 447/20. The work of J. Belikov was partly supported by the Estonian Research Council grant PRG1463. This research was partially supported by the Israel planning and budgeting committee council (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, Israel, the Technion Center for Machine Learning and Intelligent Systems (MLIS), Israel and by The Nancy and Stephen Grand Technion Energy Program (GTEP), Israel, in association with the Guy Sella Memorial Project.

References

[1] Khodayar M, Liu G, Wang J, Khodayar ME. Deep learning in power systems research: A review. CSEE J Power Energy Syst 2020;7(2):209–20.
[2] Ozcanli AK, Yaprakdal F, Baysal M. Deep learning methods and applications for electrical power systems: A comprehensive review. Int J Energy Res 2020;44(9):7136–57.
[3] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI-explainable artificial intelligence. Science Robotics 2019;4(37).
[4] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI) 6. IEEE Access 2018;52138–60.
[5] Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. 2020, [Online]. Available: arXiv:2006.11371.
[6] Ribeiro MT, Singh S, Guestrin C. Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 1135–44.
[7] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. In: NIPS. 2017, p. 1–10.
[8] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE international conference on computer vision. 2017, p. 618–26.
[9] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International conference on machine learning. PMLR; 2017, p. 3145–53.
[10] 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;29–49. http://dx.doi.org/10.1109/tkde.2020.2983930.
[11] Shi Z, Yao W, Li Z, Zeng L, Zhao Y, Zhang R, et al. Artificial intelligence techniques for stability analysis and control in smart grids: Methodologies, applications, challenges and future directions. Appl Energy 2020;278:115733. http://dx.doi.org/10.1016/j.apenergy.2020.115733.
[12] Cremer JL, Konstantelos I, Strbac G. From optimization-based machine learning to interpretable security rules for operation. IEEE Trans Power Syst 2019;34(5):3826–36.
[13] Ren C, Xu Y, Zhang R. An interpretable deep learning method for power system dynamic security assessment via tree regularization. IEEE Trans Power Syst 2021;1. http://dx.doi.org/10.1109/TPWRS.2021.3133611.
[14] Chen M, Liu Q, Chen S, Liu Y, Zhang CH, Liu R. XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access 2019;7:13149–58.
[15] Wu S, Zheng L, Hu W, Yu R, Liu B. Improved deep belief network and model interpretation method for power system transient stability assessment. J Mod Power Syst Clean Energy 2020;8(1):27–37.
[16] Han T, Chen J, Wang L, Cai Y, Wang C. Interpretation of stability assessment machine learning models based on Shapley value. In: 2019 IEEE 3rd conference on energy internet and energy system integration. 2019, p. 243–7.
[17] Kruse J, Schäfer B, Witthaut D. Revealing drivers and risks for power grid frequency stability with explainable AI. Patterns 2021;2(11):1–26.
[18] Kruse J, Schäfer B, Witthaut D. Exploring deterministic frequency deviations with explainable ai. In: 2021 IEEE international conference on communications, control, and computing technologies for smart grids. SmartGridComm, IEEE; 2021, p. 133–9.
[19] Kruse J, Schäfer B, Witthaut D. Secondary control activation analysed and predicted with explainable AI. 2021, [Online]. Available: arXiv:2109.04802.
[20] Zhang K, Xu P, Zhang J. Explainable AI in deep reinforcement learning models: A SHAP method applied in power system emergency control. In: 2020 IEEE 4th conference on energy internet and energy system integration. 2020, p. 711–6. http://dx.doi.org/10.1109/EI250167.2020.9347147.
[21] Zhang K, Zhang J, Xu P-D, Gao T, Gao DW. Explainable AI in deep reinforcement learning models for power system emergency control. IEEE Trans Comput Soc Syst 2021;1–9. http://dx.doi.org/10.1109/TCSS.2021.3096824.
[22] Santos OL, Dotta D, Wang M, Chow JH, Decker IC. Performance analysis of a DNN classifier for power system events using an interpretability method. Int J Electr Power Energy Syst 2022;136:107594. http://dx.doi.org/10.1016/j.ijepes.2021.107594.
[23] Machlev R, Perl M, Belikov J, Levy K, Levron Y. Measuring explainability and trustworthiness of power quality disturbances classifiers using XAI - explainable artificial intelligence. IEEE Trans Ind Inf 2021;1. http://dx.doi.org/10.1109/tii.2021.3126111.
[24] Zhang D, Li C, Shahidehpour M, Wu Q, Zhou B, Zhang C, et al. A bi-level machine learning method for fault diagnosis of oil-immersed transformers with feature explainability. Int J Electr Power Energy Syst 2022;134:107356.
[25] Sairam S, Srinivasan S, Marafioti G, Subathra B, Mathisen G, Bekiroglu K. Explainable incipient fault detection systems for photovoltaic panels. 2020, [Online]. Available: arXiv:2011.09843.
[26] Sairam S, Seshadhri S, Marafioti G, Srinivasan S, Mathisen G, Bekiroglu K. Edge-based explainable fault detection systems for photovoltaic panels on edge nodes. Renew Energy 2021. http://dx.doi.org/10.1016/j.renene.2021.10.063.
[27] Donti PL, Kolter JZ. Machine learning for sustainable energy systems. Ann Rev Environ Resourc 2021;46(1):719–47.

[28] Sarp S, Kuzlu M, Cali U, Elma O, Guler O. An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool. In: 2021 IEEE power energy society innovative smart grid technologies conference. 2021, p. 1–5. http://dx.doi.org/10.1109/ISGT49243.2021.9372263.

[29] Kuzlu M, Cali U, Sharma V, Güler Ö. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access 2020;8:187814–23.

[30] Chang X, Li W, Ma J, Yang T, Zomaya AY. Interpretable machine learning in sustainable edge computing: A case study of short-term photovoltaic power output prediction. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing. 2020, p. 8981–5. http://dx.doi.org/10.1109/ICASSP40776.2020.9054088.

[31] Lu Y, Murzakhanov I, Chatzivasileiadis S. Neural network interpretability for forecasting of aggregated renewable generation. In: IEEE international conference on communications, control, and computing technologies for smart grids. 2021, p. 282–8.

[32] Mitrentsis G, Lens H. An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. Appl Energy 2022;309:118473.

[33] Wang H, Cai R, Zhou B, Aziz S, Qin B, Voropai N, et al. Solar irradiance forecasting based on direct explainable neural network. Energy Convers Manage 2020;226:113487.

[34] Bahani K, Ali-Ou-Salah H, Moujabbir M, Oukarfi B. A novel interpretable model for solar radiation prediction based on adaptive fuzzy clustering and linguistic hedges. In: SITA'20: Proceedings of the 13th international conference on intelligent systems: Theories and applications. 2020, p. 1–12. http://dx.doi.org/10.1145/3419604.3419807.

[35] Chaibi M, Benghoulam EM, Tarik L, Berrada M, Hmaidi AE. An interpretable machine learning model for daily global solar radiation prediction. Energies 2021;14(21):7367.

[36] Lee YG, Oh JY, Kim G. Interpretation of load forecasting using explainable artificial intelligence techniques. Trans Korean Inst Electr Eng 2020;69(3):480–5.

[37] Alova G, Trotter PA, Money A. A machine-learning approach to predicting Africa's electricity mix based on planned power plants and their chances of success. Nat Energy 2021;6(2):158–66.

[38] Toubeau J-F, Bottieau J, Wang Y, Vallee F. Interpretable probabilistic forecasting of imbalances in renewable-dominated electricity systems. IEEE Trans Sustain Energy 2021;1. http://dx.doi.org/10.1109/TSTE.2021.3092137.

[39] Zhao S, Blaabjerg F, Wang H. An overview of artificial intelligence applications for power electronics. IEEE Trans Power Electron 2021;36:4633–58. http://dx.doi.org/10.1109/TPEL.2020.3024914.

[40] Sahoo S, Wang H, Blaabjerg F. On the explainability of black box data-driven controllers for power electronic converters. In: 2021 IEEE energy conversion congress and exposition. 2021, p. 1366–72. http://dx.doi.org/10.1109/ECCE47101.2021.9595231.

[41] Chakraborty D, Alam A, Chaudhuri S, Başağaoğlu H, Sulbaran T, Langar S. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. Appl Energy 2021;291:116807. http://dx.doi.org/10.1016/j.apenergy.2021.116807.

[42] Wenninger S, Kaymakci C, Wiethe C. Explainable long-term building energy consumption prediction using qlattice. Appl Energy 2022;308:118300.

[43] Zhang W, Liu F, Wen Y, Nee B. Toward explainable and interpretable building energy modelling: An explainable artificial intelligence approach. In: Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation. New York, NY, USA: Association for Computing Machinery; 2021, p. 255–8.

[44] Fan C, Xiao F, Yan C, Liu C, Li Z, Wang J. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. Appl Energy 2019;235:1551–60.

[45] Golizadeh Akhlaghi Y, Aslansefat K, Zhao X, Sadati S, Badiei A, Xiao X, et al. Hourly performance forecast of a dew point cooler using explainable artificial intelligence and evolutionary optimisations by 2050. Appl Energy 2021;281:116062.

[46] Kim M, Jun J-A, Song Y, Pyo CS. Explanation for building energy prediction. In: 2020 International conference on information and communication technology convergence. 2020, p. 1168–70. http://dx.doi.org/10.1109/ICTC49870.2020.9289340.

[47] Gao Y, Ruan Y. Interpretable deep learning model for building energy consumption prediction based on attention mechanism. Energy Build 2021;252:111379.

[48] Li A, Xiao F, Zhang C, Fan C. Attention-based interpretable neural network for building cooling load prediction. Appl Energy 2021;299:117238.

[49] Miller C. What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification. Energy Build 2019;199:523–36.

[50] Arjunan P, Poolla K, Miller C. Energystar++: Towards more accurate and explanatory building energy benchmarking. Appl Energy 2020;276:115413.

[51] Konstantakopoulos IC, Das HP, Barkan AR, He S, Veeravalli T, Liu H, et al. Design, benchmarking and explainability analysis of a game-theoretic framework towards energy efficiency in smart infrastructure. 2019, CoRR abs/1910.07899.

[52] Das HP, Konstantakopoulos IC, Manasawala AB, Veeravalli T, Liu H, Spanos CJ. A novel graphical Lasso based approach towards segmentation analysis in energy game-theoretic frameworks. 2019, CoRR abs/1910.02217.

[53] Bhatia A, Garg V, Haves P, Pudi V. Explainable clustering using hyper-rectangles for building energy simulation data. IOP Conf Ser: Earth Environ Sci 2019;238:012068. http://dx.doi.org/10.1088/1755-1315/238/1/012068.

[54] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

[55] Lissa P, Deane C, Schukat M, Seri F, Keane M, Barrett E. Deep reinforcement learning for home energy management system control. Energy AI 2021;3:100043. http://dx.doi.org/10.1016/j.egyai.2020.100043.

[56] Bui V-H, Hussain A, Kim H-M. Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties. IEEE Trans Smart Grid 2020;11(1):457–69. http://dx.doi.org/10.1109/TSG.2019.2924025.

[57] Yan Z, Xu Y. Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search. IEEE Trans Power Syst 2019;34(2):1653–6. http://dx.doi.org/10.1109/TPWRS.2018.2881359.

[58] Duan J, Shi D, Diao R, Li H, Wang Z, Zhang B, et al. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. IEEE Trans Power Syst 2020;35(1):814–7. http://dx.doi.org/10.1109/TPWRS.2019.2941134.

[59] Liu N, Yu X, Wang C, Li C, Ma L, Lei J. Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers. IEEE Trans Power Syst 2017;32(5):3569–83. http://dx.doi.org/10.1109/TPWRS.2017.2649558.

[60] Kim B-G, Zhang Y, Van Der Schaar M, Lee J-W. Dynamic pricing and energy consumption scheduling with reinforcement learning. IEEE Trans Smart Grid 2015;7(5):2187–98.

[61] Dalal G, Mannor S. Reinforcement learning for the unit commitment problem. In: 2015 IEEE eindhoven powertech. 2015, p. 1–6. http://dx.doi.org/10.1109/PTC.2015.7232646.

[62] Rehman AU, Lie TT, Vallès B, Tito SR. Non-invasive load-shed authentication model for demand response applications assisted by event-based non-intrusive load monitoring. Energy AI 2021;3:100055. http://dx.doi.org/10.1016/j.egyai.2021.100055.

[63] Froehlich J, Larson E, Gupta S, Cohn G, Reynolds M, Patel S. Disaggregated end-use energy sensing for the smart grid. IEEE Pervasive Comput 2011;10(1):28–39.

[64] Huber P, Calatroni A, Rumsch A, Paice A. Review on deep neural networks applied to low-frequency NILM. Energies 2021;14(9).

[65] López KL, Gagné C, Gardner M-A. Demand-side management using deep learning for smart charging of electric vehicles. IEEE Trans Smart Grid 2019;10(3):2683–91. http://dx.doi.org/10.1109/TSG.2018.2808247.

[66] Murray D, Stankovic L, Stankovic V. Explainable NILM networks. In: International workshop on non-intrusive load monitoring. 2020, p. 64–9. http://dx.doi.org/10.1145/3427771.3427855.

[67] Murray D, Stankovic L, Stankovic V. Transparent AI: Explainability of deep learning based load disaggregation. In: Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation. ACM; 2021, p. 268–71. http://dx.doi.org/10.1145/3486611.3492410.

[68] Hassan LH, Moghavvemi M, Almurib HA, Steinmayer O. Current state of neural networks applications in power system monitoring and control. Int J Electr Power Energy Syst 2013;51:134–44. http://dx.doi.org/10.1016/j.ijepes.2013.03.007.

[69] Li W, Deka D, Chertkov M, Wang M. Real-time faulted line localization and PMU placement in power systems through convolutional neural networks. IEEE Trans Power Syst 2019;34(6):4640–51. http://dx.doi.org/10.1109/TPWRS.2019.2917794.

[70] Qiu W, Tang Q, Zhu K, Wang W, Liu Y, Yao W. Detection of synchrophasor false data injection attack using feature interactive network. IEEE Trans Smart Grid 2021;12(1):659–70. http://dx.doi.org/10.1109/TSG.2020.3014311.

[71] Wang D, Wang X, Zhang Y, Jin L. Detection of power grid disturbances and cyber-attacks based on machine learning. J Inf Secur Appl 2019;46:42–52. http://dx.doi.org/10.1016/j.jisa.2019.02.008.