Interpretable Machine Learning for power systems: Establishing Confidence in SHapley Additive exPlanations

Tabia Ahmad[*]
University of Manchester, UK

Robert I. Hamilton
Shell Global Solutions International B.V., Netherlands

Jochen Stiasny
Technical University of Denmark, Lyngby

Samuel Chevalier
University of Vermont, USA

Rahul Nellikkath
Technical University of Denmark, Lyngby

Ilgiz Murzakhanov
Technical University of Denmark, Lyngby

Spyros Chatzivasileiadis
Technical University of Denmark, Lyngby

Panagiotis N. Papadopoulos
University of Manchester, UK

Abstract

Interpretable Machine Learning (IML) is expected to remove significant barriers for the application of Machine Learning (ML) algorithms in power systems. This work first seeks to showcase the benefits of SHapley Additive exPlanations (SHAP) for understanding the outcomes of ML models, which are increasingly being used to optimise power systems with increasing share of Renewable Energy (RE), to support worldwide calls for decarbonisation and climate change. To do so, we demonstrate that the Power Transfer Distribution Factors (PTDF)—a power system physics-based linear sensitivity index—can be derived from the SHAP values. To do so, we take the derivatives of SHAP values from a ML model trained to learn line-flows from generator power-injections, using a DC power-flow case in a benchmark test network. In demonstrating that SHAP values can be related back to the physics that underpin the power system, we build confidence in the explanations SHAP can offer.

1 Introduction

Tackling climate change and mitigating its effects on the ecosystem and mankind are one of the key challenges of today's world. In response to worldwide calls on decarbonisation, the Paris Agreement on climate change (2021) has set a target to minimize global warming to 1.5 °C[1]. In order to meet this ambitious target, the emission of greenhouse gases (GHGs) like CO_2 need to be drastically cut down. It is estimated that around 25% GHG emissions come from electricity and heating [2]). Therefore, in order to cut down these emissions, there is a global trend to integrate more renewable energy sources (like wind, solar, etc.) for electric power generation. To integrate them effectively into legacy power systems (both in a cost-effective and technically feasible manner), power system operators need to resort to more efficient techniques for solving day-to-day problems like optimal dispatch of generators and longer term problems like transmission system expansion planning. For example, a 2012 report from the Federal Energy Regulatory Commission (FERC) estimated that the

[*]Corresponding author – tabia.ahmad@manchester.ac.uk

Tackling Climate Change with Machine Learning: workshop at ICLR 2024.

inefficiencies induced by approximate-solution techniques may cost billions of dollars and release unnecessary emissions [3].

In recent years, many Machine Learning (ML) applications to predict the behaviour of power systems have been developed—some of which are summarised in [4]. While these black-box ML algorithms have shown good accuracy and computational savings, their applicability to mission-critical infrastructure such as power systems is limited due to absence of trustworthy explanations.

The premise of interpretable ML—an emerging area of research—is to provide detailed explanations of ML model predictions in order to enhance confidence in the model predictions. Post-hoc ML model interpretability methods are presented concisely in [5]. Notably, Local Interpretable Model-agnostic Explanations (LIME) is a local technique capable of providing feature effects for individual points that can be extrapolated to form global explanations[6]. However, a key limitation with LIME is that defining the neighbourhood around the instance for explanation is complicated and can lead to errors. Permutation Feature Importance (PFI) is a global technique[7], whereby PFI can provide feature importance, but not feature effects. The authors in [8] use gradient-based neural network interpretability methods like Integrated Gradients, Expected Gradients, and DeepLIFT for forecasting of aggregated renewable generation. SHapley Additive exPlanations (SHAP) has gained some initial traction in power systems, as [9] reviews. For example, in [10], SHAP is used to provide explanations of ML models for gaining insights for complex underlying mechanisms affecting the transient stability boundary of power systems (a phenomenon driven by complex non-linear dynamics). In [11] SHAP is used for the frequency stability problem. In [12], authors use SHAP values to determine the relationship between power system small-signal stability and topological graph metrics relating to connection of RE. The focus of the limited number of applications in literature, however, usually centre around applying SHAP, rather than analysing the method itself for proving one-to-one equivalence with the physics of the system, which is indeed non-trivial.

The SHAP [13] framework proposes using a simple linear explanation model as an interpretable approximation of the complex original model. In doing so, a class of additive feature attribution methods can be defined which unifies six existing IML methods (including LIME and DeepLIFT) demonstrating that SHAP is the only method that possesses the aforementioned properties. Such methods construct a simple additive *explanation model*, g—which is a linear function of binary variables—to represent the complex *original model*, f. In the SHAP framework, the explanation model is expressed as a "conditional expectation function of the original model" [13]. Simplified inputs x' are used to map to the original input, x through mapping function h_x, where $x = h_x(x')$; ensuring $g(z') \approx g(h_x(z'))$, whenever $z' \approx x'$ and where $z' \in \{0,1\}^M$ and M is the number of simplified input features. Thus an effect ϕ_i (where $\phi_i \in \mathbb{R}$) is attributed to each feature, the sum of which approximates $f(\boldsymbol{x})$ as per (1)

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

This work initially showcases SHAP interpretations as a tool for understanding power system ML models using a simple case study (Section 2), and then establishes the capability of SHAP explanations to capture underlying physics of the power system—thus enhancing confidence in SHAP as an interpretability method (Section 3). We achieve this by deriving PTDF (i.e., the sensitivity of line flows to power injections) from SHAP values. In doing so, we seek to build confidence in SHAP as a tool for interpreting ML models.

2 Using SHAP values to interpret ML models for power-system: Case study

Power flow analysis is used to evaluate the flow of electricity in the power system network. It is used to determine the steady-state variables, such as the voltage magnitude and phase angle of each bus and the active/reactive power flow on each line, in order to take day-to-day operational decisions. In this case study, a simple DC power problem is formulated and ML is used for line-flow prediction. Individual XGBoost [14] regression models, $f_{line,i-j}(x)$ are trained to predict the active power-flow ($F_{line,i-j}$) on each line of the the 9-bus 3-generator test network[15], as shown in Fig. 1. This is achieved using a set of input features $x \in \boldsymbol{P}$, which include generator active power injections. Since $PG1$ is the slack generator (and therefore not an independent feature), it is excluded from the features. A 75-25% train-test split is used for the XGBoost models. For the trained models, SHAP values are calculated (using the framework set out in[13], implemented in Python).

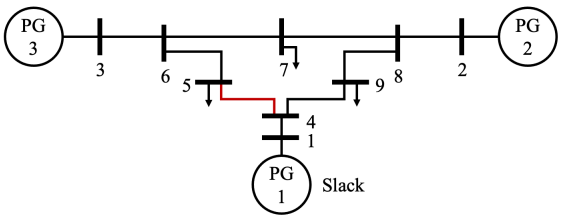
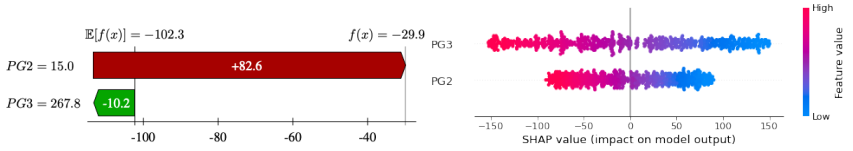


Figure 1: 9-bus 3-generator test network with line 4-5 highlighted for analysis.



(a) Local explanation for the model $f_{line,4-5}(x)$

(b) Global explanation for $F_{line,4-5}(x)$

Figure 2: Local and global SHAP explanations

2.1 Database creation

To generate the training database , samples of generator active-power injections, $PG2$ and $PG3$ respectively, are drawn independently from a uniform distribution in the range [0, 500] MW for a total of 1001 generation scenarios. For the sake of this case study, but without loss of generality, demand is assumed constant at 315 MW. Solving DC power flow, the active power flow on each line is recorded ($F_{line,i-j}$) in the database. The final database contains $PG2$ and $PG3$ as training features and $F_{line,i-j}$ as targets for all 1001 operational scenarios, available at [16]. To gain insights into each ML model, the Tree SHAP algorithm [17] was used. Analysis of a single local explanation and a global model interpretation of $f_{line,4-5}(x)$ is presented below. Similar analysis can be extracted for the remaining ML models $f_{line,i-j}(x)$.

2.1.1 Results: Local Explanation

Local explanations provide the contributions of features (i.e. SHAP) values, ϕ_i) in shifting the model prediction from the model expectation ($\mathbb{E}[f(x)]$)—that would be predicted if we had no feature information—to the final prediction ($f(x)$) for a single operational scenario. Therefore the SHAP values (ϕ_i) are given in the units of the variable being predicted. In this study, features are the $PG2$ and $PG3$ injections. An example of a local SHAP explanation for the model $f_{line,4-5}(x)$ is given in Fig. 2(a). For this particular operational scenario, the baseline model expectation ($\mathbb{E}[f(x)]$) for $F_{line,4-5}$ is -102.3 MW. $PG3$ setpoint is 267.8 MW, which decreases the prediction by -10.2 MW (ϕ_{PG3})) from $\mathbb{E}[f(x)]$ to -112.5 MW. The setpoint for $PG2$ is 15.0 MW, which increases the model prediction by 82.6 MW (ϕ_{PG2}). This results in the final prediction of $f(x) = -29.9$ MW. This is consistent with (1). Since $|\phi_{PG2}| > |\phi_{PG3}|$, $PG2$ has a greater effect in this scenario and thus is placed higher on the y-axis.

2.1.2 Results: Global Interpretation

Global interpretations comprise of the local explanations for the entire training database, making them consistent. Analysis of SHAP values in the global frame assists in understanding the global model structure. The global SHAP plot (Fig. 2(b)) gives the SHAP value for each feature (x-axis) with respect to the feature value (color-axis). Features are ordered on the y-axis based on importance (defined here as the mean of all SHAP values for all operational scenarios). In this case, $PG3$ is found to have a higher importance than $PG2$ (note: this is the inverse to the local explanation given in Fig. 2(a) for that particular case, indicating how local sensitivity vs. global importance need not necessarily be the same). It can be seen that as $PG3$ increases, the SHAP values (ϕ_{PG3}) decrease from positive to negative. This means that as $PG3$ increases the impact it has on $F_{line,4-5}$ goes from increasing the baseline prediction, to decreasing it. This is consistent with the theory of

power flow when observing Fig. 1—where one would expect a higher $PG3$ setpoint to decrease $F_{line,4-5}$ in this simple DC power flow example. A similar trend can be seen for the SHAP values for $PG2$ (ϕ_{PG2})—although the impact is smaller than $PG3$, indicated by a smaller spread of SHAP values on the x-axis. The feature value is given on the color axis and is normalised based on the min/maximum feature value for trend identification, however the actual feature value (in MW) can also be extracted. This showcases how SHAP values can give some interesting insights about how power system features, here the power injections, affect our outputs of interest, here the line flows. This can be especially useful in more complex cases of RE-integrated power systems, where analytical relationships are difficult to extract.

3 Derivation of PTDF from SHAP values

To show the link between SHAP values and the PTDF analytically, we consider $f_{line,i-j}(x)$ to be a linear statistical model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ whose features ($\boldsymbol{x} \in \boldsymbol{P}$) are assumed to be independent. Here the features are the power injections $PG2$ and $PG3$.

Theorem 1. *The derivatives of the SHAP values $\phi_i(f, \boldsymbol{x})$, $i \neq 0$, associated with $f(\boldsymbol{x})$ yield exactly the PTDF of this network.*

Proof. Using Corollary 1 from [13], the SHAP values $\phi_i(f, \boldsymbol{x})$ associated with $f(\boldsymbol{x})$ are given by

$$\phi_0(f, \boldsymbol{x}) = b \tag{2}$$

$$\phi_i(f, \boldsymbol{x}) = \boldsymbol{w}_i(\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{X}_i]),\ i \neq 0 \tag{3}$$

where $\boldsymbol{X}_i$ is the training data associated with the i^{th} feature. Perturbing the i^{th} feature (i.e., continuous regressor) yields

$$\frac{\partial \phi_i}{\partial \boldsymbol{x}_i} = \boldsymbol{w}_i. \tag{4}$$

Since $\boldsymbol{w}_i$ relates the sensitivity of line flow to the i^{th} injection, the SHAP derivative is equivalent to a PTDF. □

For a definition and analytical derivation of the PTDF, the interested reader can refer to [18]. Strictly speaking, the result of Theorem 1 is only valid when the underlying statistical model is linear (or affine). However, ML practitioners often use models which have the capacity to be aggressively *nonlinear*. SHAP can be applied in either case and as the trained models effectively still behave like *linear* models, (3)-(4) will remain approximately valid. To show this experimentally, we note that the sum across all SHAP values should yield a model with linear sensitivity to power injections, i.e., an approximate PTDF vector $\hat{\boldsymbol{D}}$, and a constant offset term ϵ:

$$\sum_i \phi_i(f, x) = \boldsymbol{w}^\top(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}]) + b \approx \hat{\boldsymbol{D}}^\top \boldsymbol{P} + \epsilon. \tag{5}$$

By collecting a library of SHAP values Φ associated with a library of sampled injection values $\mathbf{P}$, a regression procedure can yield the PTDF approximation $\hat{\boldsymbol{D}}$:

$$\begin{bmatrix} \hat{\boldsymbol{D}} \\ \epsilon \end{bmatrix} = \mathbf{P}^{+}\Phi, \tag{6}$$

where $(\cdot)^+$ denotes Moore–Penrose pseudoinversion. Table 1 presents the true physical PTDF and the SHAP-based PTDF $\hat{\boldsymbol{D}}$ for the relevant buses for the case study in Section II. These experimental results support Theorem 1, showing that the derivatives of the SHAP values are equivalent to the PTDF for this network.

4 Conclusion

Interpretable ML is expected to play an important role in understanding increasingly complex models, necessary for operating decarbonised power-systems, on the pathway to mitigate climate change.

Table 1: 9-Bus 3-Generator Test Network PTDF & SHAP.

	True physical PTDF, $\boldsymbol{D}$		SHAP-based PTDF, $\hat{\boldsymbol{D}}$	
Line	Bus 2	Bus 3	Bus 2	Bus 3
Line 1-4	-1.0000	-1.0000	-0.9999	-0.9999
Line 4-5	-0.3613	-0.6152	-0.3613	-0.6151
Line 5-6	-0.3613	-0.6152	-0.3613	-0.6151
Line 3-6	0	1.0000	0.0000	0.9999
Line 6-7	-0.3613	0.3848	-0.3613	0.3848
Line 7-8	-0.3613	0.3848	-0.3613	0.3848
Line 8-2	-1.0000	0	-1.0000	0.0000
Line 8-9	0.6387	0.3848	0.6386	0.3848
Line 9-4	0.6387	0.3848	0.6386	0.3848

In this direction, the paper establishes one-to-one correspondence between SHAP interpretations for power system ML models and power system physics. For wide-spread adoption of ML models developed for safely and reliably operating decarbonised power systems, confidence must be built in the interpretation method. Physical equivalence is one such way of developing confidence, which we have shown here for SHAP and PTDF in a simple DC power flow case. Extending from this linear case to more complex non-linear problems will likely be a promising avenue of future research.

Acknowledgments and Disclosure of Funding

T. Ahmad, R.I. Hamilton and P. N. Papadopoulos are supported by the UKRI Future Leaders Fellowship MR/S034420/1. For the purpose of open access, the authors have applied for a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission. All results can be fully reproduced using the methods and data described in this paper and provided references. Jochen Stiasny, Rahul Nellikkath, Ilgiz Murzakhanov and Spyros Chatzivasileiadis are supported by the ID-EDGe project, funded by Innovation Fund Denmark, Grant Agreement No. 8127-00017B, and by the ERC Starting Grant VeriPhIED, Grant Agreement No. 949899. Samuel Chevalier is supported by the HORIZON-MSCA-2021 Postdoctoral Fellowship Program, Project #101066991 – TRUST-ML.

References

[1] Ringo Doe. Goals of UKCOP 26 conference on climate change. `https://ukcop26.org/cop26-goals/`, May 2021.

[2] Global greenhouse gases emission data. `https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data`, May 2021.

[3] Mary B Cain, Richard P O'neill, Anya Castillo, et al. History of optimal power flow and formulations. *Federal Energy Regulatory Commission*, 1:1–36, 2012.

[4] Asiye K Ozcanli, Fatma Yaprakdal, and Mustafa Baysal. Deep learning methods and applications for electrical power systems: A comprehensive review. *International Journal of Energy Research*, 44(9):7136–7157, 2020.

[5] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, pages 1135–1144, 2016.

[7] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[8] Yucun Lu, Ilgiz Murzakhanov, and Spyros Chatzivasileiadis. Neural network interpretability for forecasting of aggregated renewable generation. In *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 282–288. IEEE, 2021.

[9] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9:100169, 2022.

[10] Robert I Hamilton and Panagiotis N Papadopoulos. Using SHAP values and machine learning to understand trends in the transient stability limit. *IEEE Trans. on Power Syst.*, 2023.

[11] Johannes Kruse, Benjamin Schäfer, and Dirk Witthaut. Revealing drivers and risks for power grid frequency stability with explainable AI. *Patterns*, 2(11), 2021.

[12] Wenting Yi and David J Hill. Topological stability analysis of high renewable penetrated systems using graph metrics. In *2021 IEEE Madrid PowerTech*, pages 1–6. IEEE, 2021.

[13] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] Tianqi Chen and Carlos Guestrin. XGBoost: a scalable tree boosting system ACM SIGKDD international conference on knowledge discovery and data mining. *ACM*, pages 785–794, 2016.

[15] Ray D Zimmerman, Carlos E Murillo-Sánchez, and Deqiang Gan. MATPOWER. *PSERC.[Online]. Software Available at: http://www. pserc. cornell. edu/matpower*, 1997.

[16] SHAP Database Repository. `https://github.com/an7ubr6498/SHAP-database`, 2022.

[17] Scott M Lundberg et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[18] Spyros Chatzivasileiadis. Lecture notes on Optimal Power Flow (OPF). *CoRR*, abs/1811.00943, 2018.