TimeXAI: Unified Archive and Concept-Based Counterfactual Explanations for Sustainable Energy Time Series

Anonymous submission

Abstract

Explaining AI systems operating on time series data is crucial in many decision-making areas, such as healthcare, energy, and public policy-making, which requires interpretable and transparent explanations to overcome the black-box nature of models, especially for non-experts. Effective explanations allow us to understand how a model has learned, which helps in taking steps to improve robustness, safety, and fairness. Concept-based explanations have gained traction, offering insights into AI decisions using higher-level concepts. Concurrently, in our climate-conscious world, businesses increasingly rely on time series data to enhance energy efficiency and drive sustainable practices. Yet, several significant challenges persist. There is a lack of comprehensive archives for sustainable energy time series data, and current models often lack robust, regression-explainable methods to interpret their behavior. Our findings indicate that many existing models are prone to over-fitting specific open-source datasets, resulting in a disconnect between their performance in controlled environments and real-world applications. To address this, we introduced TIMEXAI, a framework that uses counterfactual-based explanations to uncover these weaknesses and provide deeper insights into where and why models struggle. To further this effort, we introduce a comprehensive archive of 78 publicly available sustainable energy time series datasets and a newly collected dataset, encompassing a total of over 137 million hourly instances at a 1Hz sampling rate. Our results strongly suggest that future work should explore more varied set time series to better assess model performance and prevent the risk of over-fitting to specific time series data sets. The archive and code can be accessed at https://TimeXAI.github.io/.

Introduction

Deep learning has become a cornerstone technology in analyzing energy time series data, prevalent in scenarios such as electricity consumption prices, household energy prediction, and flexibility in the smart grid. Despite its successes, a critical limitation of deep learning in time series analysis is the lack of explainability, which is crucial for gaining trust and actionable insights in these sensitive and impactful domains. Given the rising electricity consumption prices and the increasing need for sustainable practices, society is actively searching for ways to change user behavior. This mission is fueled by rising electricity prices and the urgency of adopting eco-friendly practices. Current efforts in enhancing explainability for time series analysis primarily focus on pinpointing

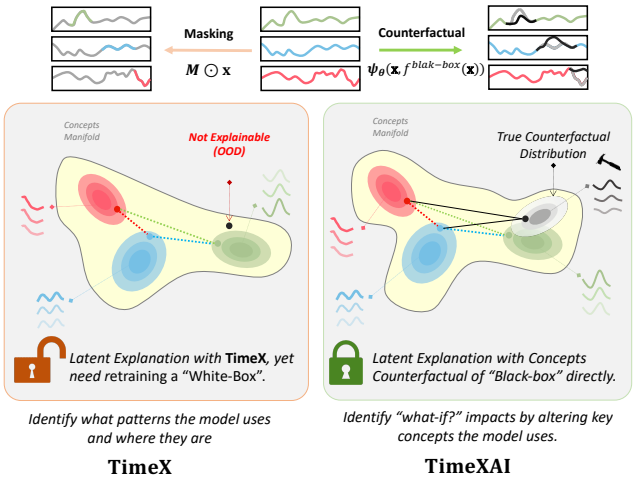

Figure 1: Comparison of our TIMEXAI with TIMEX (Queen et al. 2023). TIMEXAI directly works with the black-box model, slightly modifying the original time series $\mathbf{x}$ to handle out-of-distribution (OOD) cases and find counterfactuals. In contrast, TIMEX builds a white-box model to approximate the black-box behavior, which may cause OOD scenarios.

significant locations of time series signals that dominate the model's prediction in a post-hoc sense. For example, (Shi, Stebliankin, and Narasimhan 2023) explained their trained models for water level prediction using LIME, a local interpretable model-agnostic explanation technique (Ribeiro, Singh, and Guestrin 2016a). On top of this intuitive principle, perturbation-based methods, including DYNAMASK (Crabbé and Van Der Schaar 2021) and Extrmask (Enguehard 2023), offer insights by altering non-salient features to assess their impact on model output. However, without a theoretical foundation, these ad-hoc designed objectives are often specific to a single domain and do not generalize to wider scenarios. One area of focus is how household electricity usage is visually presented, aiming to create a smarter and more conscious approach. Energy time series data stands apart with distinct features. These include noticeable seasonal patterns, recurring sequences over days and seasons, and the predictable daily rhythm characterized by surges during active hours and lulls at night. The shift between weekdays and weekends adds further variability, sometimes punctuated by abrupt surges linked to external influences. Irregular consumption and de-

viations from routines contribute to complexity. The intricate interplay of factors such as weather and household occupancy adds to this, not to mention the diverse energy types shaping consumption patterns.

Challenge 1 (WHAT-IF? Counterfactual Explanation). *Consider a black-box model $f(\cdot)$ that takes as input a time series of energy consumption $\mathbf{x} \in \mathbb{R}^T$ from a household within a smart grid system to predict the corresponding energy demand category as 'High Demand'. Traditional methods might assign importance scores to each entry of $\mathbf{x}$ to explain this prediction. However, they often fail to provide* actionable *counterfactual explanations on how to adjust $\mathbf{x}$ to $\tilde{\mathbf{x}}$ such that the prediction could shift to $\tilde{y}$ that is 'Low Demand'. Such a method would be valuable in understanding how specific changes in energy consumption patterns could lead to different demand outcomes, thus offering deeper insights into the model's decision-making process.*

Fortunately, forecasting and energy disaggregation have revolutionized electricity monitoring by leveraging deep learning algorithms. This enables analyzing consumption patterns without invasive measurements or appliance sensors. Beyond individual appliance identification and detailed energy breakdown, offering critical insights for energy planning and management. This empowers companies to forecast electricity demands, allocate resources efficiently, optimize energy use, reduce costs, and contribute to sustainability. This progress has spurred many studies (Zhang et al. 2021a; Letzgus et al. 2021; Siddiqui et al. 2019). However, constrained by limited datasets in terms of variety and available attributes, many methods lean exclusively on power or voltage, leading to less effective real-world outcomes.

Thus, it is extremely difficult for researchers to fairly compare models and develop new models based on only thus limited datasets. To fill this gap, we introduced a novel household dataset for fine-grained designed for evaluating global and multivariate forecasting and disaggregation models. Furthermore, we develop an open-source archive and toolkit for benchmarking, referred to as TIMEXAI time series for energy, which includes a variety of generative models as shown in Fig.1, to extend appliance signatures and handle instances of reduced activation through imputation. Additionally, evaluating generated time series is still a challenge due to the lack of accessible user-friendly evaluation tools and public benchmarks to make the research results more rigorous, and reproducible. Our TIMEXAI framework consists of three core components.

The binary mask method for generating explanations in time series data can lead to limitations in interpretability. By applying a fixed binary mask to isolate sub-instances, this approach may lose important contextual information and fail to capture complex feature interactions. Additionally, it can produce trivial explanations that do not provide meaningful insights into the model's decision-making process.

Challenge 2 (In- and Out-of-Distribution Counterfactual Explanations for Continuous Predictions). *As in challenge 1, consider a time series dataset $\mathbf{x} \in \mathbb{R}^{K \times T}$ representing hourly energy consumption and variables data from a building, with a model predicting the future energy demand as $\hat{\mathbf{y}} \in \mathbb{R}^T$. To generate a counterfactual explanation that changes $\hat{\mathbf{y}}$, typical methods might suggest altering the input features $\mathbf{x}$ in ways that are unrealistic, such as suggesting changes in factors like weather conditions that are beyond the model's control. An ideal approach would involve identifying realistic changes that can be applied to the energy consumption data while keeping external factors like weather or occupancy levels unchanged. For instance, a realistic counterfactual explanation $\tilde{\mathbf{x}}$ would involve adjustments to energy usage patterns that could lower the predicted demand, while remaining within practical operational constraints.*

In challenge 2, generating realistic counterfactual explanations highlights a common issue with traditional methods. These methods might suggest impractical changes, such as altering weather conditions, which are often out-of-distribution (OOD) for the model as shown in Fig.1. This issue mirrors the limitations of the binary mask approach, which can lead to explanations that are disconnected from practical realities. In contrast, a more robust counterfactual method focuses on feasible adjustments, like modifying energy usage patterns while keeping external variables unchanged, ensuring that explanations are both realistic and actionable.

- We examine the limitations of current explanation models for time series learning through the lens of information theory and propose a practical objective function.
- We introduce a new Comprehensive Archive, serving as a downstream evaluation suite, which includes **78** datasets and features our newly collected dataset from French households (see Appendix E).
- We have developed an automated toolkit for hyperparameter tuning that supports the training of generative models and streamlines model evaluation for downstream tasks.
- We employ a diverse set of evaluation metrics to assess sample and parameter efficiency. For a visual representation of the TIMEXAI framework, please refer to Appendix B.

Paper Organization. We begin with an overview of the background and related work, followed by a detailed description of the proposed methods. Next, we present the unified archive and provide comprehensive details on the newly introduced dataset. Additional information about the policy of collected data is available in the Appendix E.

Notations and Preliminary

Notations and Problem Formulation. We focus on generating counterfactual explanations for predictions from time series models. We assume the model takes as input a multivariate time series $\mathbf{x}_i \in \mathbb{R}^{C \times T}$, where T is the length of the time series, and C is the feature dimension, and predicts the corresponding label $\mathbf{y}_i$, which can be a categorical label, a real value, or a univariate $\mathbf{y}_i \in \mathbb{R}^{1 \times T}$ or multivariate time series $\mathbf{y}_i \in \mathbb{R}^{C \times T}$. Given a specific input $\mathbf{x}_i$ and the black-box model's prediction $\mathbf{y}_i^{pred}$, our goal is to explain the model by finding a counterfactual time series $\mathbf{x}_i^{cf} \neq \mathbf{x}_i$ that could have lead the model to an alternative (counterfactual) prediction $\mathbf{y}_i^{cf}$. The value of the feature indexed c at time t is denoted

by $\mathbf{x}[t, c]$. A training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)|i \in [N]\}$ consists of N time series instances $\mathbf{x}_i$ along with their associated labels $\mathbf{y}_i$, where $\mathbf{y}_i \in \mathcal{Y}$. To develop a generally applicable model for explainability, we consider explanation methods which are task-agnostic and treat the to-be-explained model $f(\cdot)$ as a black box, i.e., the so-called *post-hoc, instance-level* explanation methods (Zhang et al. 2021b). In this context, an explanation refers to a sub-instance of the input time series, extracted using a saliency mask, which is a *sufficient statistic* of the input concerning its label.

As can be observed in the problem statement, in order to find good post-hoc instance-level time series explanations, given an observed instance X, one needs to optimize the choice of binary mask $M \in \{0, 1\}^{T \times D}$ with respect to an underlying objective function, e.g., the information bottleneck objective function discussed in the subsequent sections. In this work, we transform this discrete optimization problem into a continuous one, and consider stochastic masks. That is, we define an explanation extractor $g(\cdot)$ as a function that takes the instance X as input, and outputs a matrix $\pi = [\pi_{t,d}]_{t \in [T], d \in [D]} \in [0, 1]^{T \times D}$. Then, the binary mask is generated by producing each $M[t, d]$ independently and according to a Bernoulli distribution with parameter $\pi_{t,d}$.

Background and Related Work

Interpretation Methods for Neural Networks. Various attribution-based interpretation methods have been proposed in recent years. Some methods focused on local interpretation (Ribeiro, Singh, and Guestrin 2016b; Lundberg and Lee 2017a; Plumb, Molitor, and Talwalkar 2018; Chen et al. 2018a; Wang, Zhou, and Bilmes 2019) while others are designed for global interpretation (Ghorbani et al. 2019; Natesan Ramamurthy et al. 2020). The main idea is to assign attribution, or importance scores, to the input features in terms of their impact on the prediction (output). For example, such importance scores can be computed using gradients of the prediction with respect to the input (Selvaraju et al. 2017; Lundberg and Lee 2017b; Shrikumar, Greenside, and Kundaje 2017; Sundararajan, Taly, and Yan 2017). Some interpretation methods are specialized for time series data; these include perturbation-based (Pan, Hu, and Chen 2021), rule-based (Rajapaksha and Bergmeir 2022), and attention-based methods (Heo et al. 2018; Lim et al. 2021). One typical method, Feature Importance in Time (`FIT`), evaluates the importance of the input data based on the temporal distribution shift and unexplained distribution shift (Tonekaboni et al. 2020). However, these methods can only produce importance scores of the input features for the current prediction and therefore cannot generate counterfactual explanations.

Counterfactual Explanations for Time Series Models. There also works that generate counterfactual explanations for time series models. (Dhaou et al. 2021) proposed an association-rule algorithm to explain time series prediction by finding the frequent pairs of timestamps and generating counterfactual examples. (Nemirovsky et al. 2022) proposed a general explanation framework that generates counterfactual examples using residual generative adversarial networks (RGAN); it can be adapted for time series models. However, these works either fail to generate realistic counterfactual explanations (due to discretization error) or fail to generate feasible counterfactual explanations for time series models. In contrast, our TimeXAI as a principled variational causal method (Wang et al. 2020; Mao et al. 2021b; Gupta et al. 2021) can naturally generate realistic and feasible counterfactual explanations. Such advantages are empirically verified in .

Bayesian Deep Learning and Variational Autoencoders. Our work is also related to the broad categories of variational autoencoders (VAEs) (Sirojan, Phung, and Ambikairajah 2018) (which use inference networks to approximate posterior distributions) and Bayesian deep learning (BDL) (Wang, Wang, and Yeung 2015; Wang and Yeung 2016; Wang 2017; Huang, Wang, and Mak 2019; Wang et al. 2019; Wang and Yeung 2020; Ding et al. 2022) models (which use a deep component to process high-dimensional signals and a task-specific/graphical component to handle conditional/causal dependencies). (Lin et al. 2022) proposed the first VAE-based model for generating causal explanations for graph neural networks. (Louizos et al. 2017; Pawlowski, Coelho de Castro, and Glocker 2020) proposed the first VAE-based models for performing causal inference and estimating treatment effect. However, none of them addressed the problem of counterfactual explanation, which involves solving an inverse problem to obtain the optimal counterfactual input. In contrast, our TimeXAI is the first VAE-based model to address this challenge, with theoretical guarantees and promising empirical results. From the perspective of BDL (Wang and Yeung 2016, 2020), TimeXAI uses deep neural networks to process high-dimension signals (i.e., the deep component in (Wang and Yeung 2016)) and uses a Bayesian network to handle the conditional/causal dependencies among variables (i.e., the task-specific or graphical component in (Wang and Yeung 2016)). Therefore, TimeXAI is also the first BDL model for generating counterfactual explanations.

A Look at Existing Datasets and Data Augmentation. Recently, generative models have excelled at producing synthetic data that closely resembles authentic data, making them invaluable for tasks like data augmentation, imputation (Jeha et al. 2021), scenario simulation, and style transformation. Three main strategies are employed for generating time-series data. Auto-regressive models decompose sequence distributions into conditional probabilities, modeling them through maximum likelihood principles. While effective for stepwise forecasting, ARNs might lack diversity in sequence generation. Auto-encoders(AEs) (Demir et al. 2021) and Variational Auto-encoders (Vahdat and Kautz 2020) use an encoder-decoder architecture for sequential generation. However, AEs heavily rely on existing patterns, potentially constraining variety. Generative Adversarial Networks (GANs) (Donahue, McAuley, and Puckette 2018; Jeha et al. 2021; Yoon, Jarrett, and Van der Schaar 2019; Wang, Yan, and Oates 2017) aim to map the distribution of real data sequences to random noise, allowing flexible sampling and diverse sequence generation. GANs incorporate extra information to guide their generator and discriminator modules, enabling effective learning of joint probability distributions. Prior works like RCGAN (Esteban, Hyland, and Rätsch 2017), WaveGAN (Yamamoto,

Song, and Kim 2020), and Continuous RNN-GAN (C-RNN-GAN (Mogren 2016)) (Mogren 2016) integrate both temporal dynamics and static features into generation. While studies like (Yoon, Jarrett, and Van der Schaar 2019) have formulated GAN-based time-series generation with enhanced temporal dynamics, these often use custom evaluation settings. However, there's a lack of benchmarks to assess generative data in energy time series scenarios.

Concept-Based Counterfactual Explanation

We generate counterfactual explanations via counterfactual inference. Our goal is to find the optimal counterfactual explanation $\tilde{\mathbf{x}}$ defined below.

Definition 1 (Optimal Counterfactual Explanation). Given a factual observation $\mathbf{x}$ and prediction $\hat{\mathbf{y}}$, the optimal counterfactual explanation $\tilde{\mathbf{x}}$ for the counterfactual outcome for $\tilde{\mathbf{y}}$ is

$$\tilde{\mathbf{x}} = \arg\max_{\mathbf{x}'} \; p(Y_{\mathbf{x}=\mathbf{x}'}(\mathbf{c}) = \tilde{\mathbf{y}}),$$

where $\mathbf{c} = (\mathbf{c})$ and the counterfactual likelihood is defined as

$$p(Y_{\mathbf{x}=\mathbf{x}'}(\mathbf{c}) = \tilde{\mathbf{y}}) = \sum_{\mathbf{c}} p\left(\mathbf{y} = \tilde{\mathbf{y}} | do\left(\mathbf{x} = \mathbf{x}'\right), \mathbf{c}\right) p(\mathbf{c} | \mathbf{x} = \mathbf{x}, \mathbf{y} = \hat{\mathbf{y}}).$$

In words, we search for the optimal $\tilde{\mathbf{x}}$ that would have shifted the model prediction from $\hat{\mathbf{y}}$ to $\tilde{\mathbf{y}}$ while keeping $(\mathbf{c})$ unchanged. Since the definition of counterfactual explanations in def:counter involves causal inference with the intervention on $\mathbf{x}$, we need to first *identify* the causal probability $p(\mathbf{y} = \tilde{\mathbf{y}} | do(\mathbf{x} = \mathbf{x}'), \mathbf{c})$ using observational probability, i.e., removing 'do' in the equation. The theorem below shows that this is achievable.

Theorem 2. *Let $\mathbf{c} = \{c_1, \ldots, c_m\}$ represent a set of concepts, and let $p(\mathbf{c}|\mathbf{x}, \mathbf{y})$ denote the posterior distribution of the concept variables. The effect of an action, denoted by $p(\mathbf{y} = \tilde{\mathbf{y}} | do(\mathbf{x} = \mathbf{x}'), \mathbf{c})$, can be identified by the expectation of the posterior distribution over the concepts $\mathbb{E}_{p(\mathbf{c}|\mathbf{x}')}\left[p\left(\tilde{\mathbf{y}}|\mathbf{c}\right)\right]$.*

We provide the complete proof in Appendix A. We can now rewrite our equation as:

$$\mathcal{L}_{cf} = _{p(c_1,\ldots,c_m|\mathbf{x}=\mathbf{x},\mathbf{y}=\hat{\mathbf{y}})} \; _{p(\mathbf{z}|\mathbf{x}')}[p(\tilde{\mathbf{y}}|\mathbf{z})], \qquad (0.1)$$

where $\mathbf{c} = (c_1, \ldots, c_m)$ and $p(\mathbf{c}|\mathbf{x} = \mathbf{x}, \mathbf{y} = \hat{\mathbf{y}})$ is approximated by $q_\phi(\mathbf{c}|\mathbf{x} = \mathbf{x}, \mathbf{y} = \hat{\mathbf{y}})$. We use Monte Carlo estimates to compute the expectation, iteratively compute the gradient $\frac{\partial \mathcal{L}_{cf}}{\partial \mathbf{x}'}$ (via back-propagation) to search for the optimal $\mathbf{x}'$ in a way similar to (Wang et al. 2019; Mao et al. 2021a), and use it as $\tilde{\mathbf{x}}$ (see the complete algorithm in Appendix A).

A New Archive with Data Augmentation

In our proposed dataset, we employed a specialized SmartPlug, to facilitate a comprehensive approach to energy data collection and analysis. These Smart Plugs offer a wide range of functionalities that enable us to not only monitor the energy consumption patterns of various electrical devices but also exercise control over their operation. This ability to interact with devices at the plug level is crucial for obtaining granular insights into energy consumption behaviors. The SmartPlug serves as a multifunctional tool with several key applications and features. Firstly, it enables real-time energy consumption monitoring of connected devices, allowing us to track their power usage over time. This information is valuable for understanding the energy usage patterns and identifying potential efficiency improvements. Moreover, the SmartPlug provides a means for remotely controlling electrical devices, facilitating energy conservation strategies by allowing users to switch off devices when not in use, even when away from home. Further details on the policy for the collected data are provided. Additionally, all relevant sustainable energy data can be accessed in the TIMEXAI-ARCHIVE module, designed for user-friendly interaction.

```
from TimeXAI import load_dataset
dataset = load_dataset("Archive/Etth2",
        split="train")
```

As alluded earlier, generative Models offer an approach to enhance energy time series datasets through augmentation, imputation, and addressing missing value gaps. Evaluating these models is crucial, as we tackle challenges related to fidelity, usability and diversity.

Generative Adversarial Networks. Consider the general data set where each instance consists of two elements: static and temporal features (that occur over time, e.g., vital signs). Let $\mathcal{S}$ be a vector space of static features, $\mathcal{X}$ of temporal features, where $\mathbf{s} \in \mathcal{S}$, and $\mathbf{x} \in \mathcal{X}$ be random vectors that can be instantiated with specific values at each time step $t \in T$. We consider tuples of the form $(\mathbf{s}, \mathbf{x})$ with some joint distribution p. The goal is to train discriminator D to learn a density $\hat{p}(\mathbf{s}, \mathbf{x})$ that best approximates $p(\mathbf{s}, \mathbf{x})$. This is a high-level objective and depending on the lengths, dimensionality, and distribution of the data may be difficult to optimize in the standard GAN framework. Therefore, as stated in (Yoon, Jarrett, and Van der Schaar 2019; Shi, Srey, and Tsang; Jeha et al. 2021) inspired by the autoregressive, the joint can decomposition as $p(\mathbf{s}, \mathbf{x}) = p(\mathbf{s}) \prod_{t=1}^{T} p(\mathbf{x}_t | \mathbf{s}, \mathbf{x}_{1:t-1})$ to focus specifically on the conditionals, yielding the complementary and simpler objective of learning a density $\hat{p}(\mathbf{x}_t | \mathbf{s}, X_{1:t-1})$ that best approximates $p(\mathbf{x}_t | \mathbf{s}, \mathbf{x}_{1:t-1})$ at any time t. Importantly, this breaks down the sequence-level objective (matching the joint distribution) into a series of stepwise objectives (matching the conditionals). Global objective: $\min_{\hat{p}} D(p(\mathbf{s}, \mathbf{x}) || \hat{p}(\mathbf{s}, \mathbf{x}))$, and local objective $\min_{\hat{p}} D(p(\mathbf{x}_t | \mathbf{s}, \mathbf{x}_{t-1}) || \hat{p}(\mathbf{x}_t | \mathbf{s}, \mathbf{x}_{t-1}))$. Using a real set for supervision via maximum likelihood training, the latter takes the form of the Kullback-Leibler divergence. The objective, then, will be a combination of the GAN objective global and local, this is common for Progressive Self Attention GANs for Synthetic Time Series (PSA-GAN)(Jeha et al. 2021), WaveGAN(Yamamoto, Song, and Kim 2020), GTGAN (Jeon et al. 2022), and TimeGAN (Yoon, Jarrett, and Van der Schaar 2019). For an in-depth theoretical state, refer to Appendix C.

What defines good generated data for sustainable energy? Good generated data for energy forecasting and dis-

aggregation should possess several key characteristics from an energy perspective. These characteristics include fidelity, usability, and variety.

1. **Fidelity** assesses the ability of generated data to replicate essential characteristics of the real energy data, making them indistinguishable from real data. This is quantitatively evaluated with a discriminative score, which gauges a classifier's accuracy in distinguishing real from synthetic data. A low score indicates faithful representation of real properties.
2. **Usability** involves maintaining practical and predictive qualities in synthetic data. Generated data should support forecasting and disaggregation, retaining predictive capabilities of the real energy data. Usability is assessed with a predictive score derived by training a predictor on synthetic data and measuring MAE and RMSE. A low score signifies capture of real data's predictive traits.
3. **Explainability.** Our framework not only includes metrics to evaluate the explanation methods but also benchmarks various other methods (e.g., LIME (Lim et al. 2019), Gradient-based methods (Simonyan, Vedaldi, and Zisserman 2013)). Notably, these methods are easily accessible via the TimeXAI toolkit. The code below demonstrates DeepAR's explaining performance using GRAD (Simonyan, Vedaldi, and Zisserman 2013) metric on real data, GTGAN, and TIMEGAN.

```
from TimeXAI import Explainer
explain=Explainer(real_x=X,
    gen_model=["TimeVAE"],
    black_box_model="DeepAR",
    method="TimeXAI", metric="CCR")
```

Experiments and Use Case

In this section, we evaluate the performance of our TimeXAI method alongside existing approaches using generated and real-world datasets available in our proposed archive. For each dataset, we assess different methods based on three key metrics: (1) **prediction accuracy**, (2) **counterfactual accuracy**, and (3) **counterfactual change ratio**, with the latter being the most critical metric. The specific implementation of these metrics varies across datasets (see details in Appendix E).

Baselines and Implementations

We compare TimeXAI with several leading methods for generating explanations for deep learning models, including Regularized Gradient Descent (RGD) (Wachter, Mittelstadt, and Russell 2017), Gradient-weighted Class Activation Mapping (GRADCAM) (Selvaraju et al. 2017), Gradient SHapley Additive exPlanations (GRADSHAP) (Lundberg and Lee 2017b), Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016b), Feature Importance in Time (FIT) (Tonekaboni et al. 2020), Case-crossover APriori (CAP) (Dhaou et al. 2021), and Counterfactual Residual Generative Adversarial Network (COUNTERGAN) (Nemirovsky et al. 2022) (see **??** for further details). Notably, among these baselines, only RGD and CounteRGAN can produce actionable explanations. Other methods, such as FIT (designed specifically for time series models), only provide importance scores as explanations; therefore, some evaluation metrics may not apply to them (indicated as '-' in tables). To ensure a fair comparison, the prediction model in all baseline explanation methods uses the same neural network architecture as the inference module in our TimeXAI. Further details on the architecture, training, and inference can be found in the Appendix. Partial results are presented here due to space limitations; additional results and scripts are available in the Supplementary Materials.

Table 1: **Average CCR for intra-dataset and cross-dataset settings.** The average CCR is computed across 6 counterfactual action settings (Low→High, High→Low, Low→Low+, High→High+, High→High) for each model and dataset. The *last row* displays the average CCR over all 6 dataset settings for each method.

	CounteRGAN	RGD	GradCAM	GradSHAP	LIME	FIT	CAP	TimeXAI (Ours)
→Dataport	1.148	1.140	1.125	0.910	0.928	1.065	1.028	**1.320**
→ETTh1	1.220	1.245	1.060	0.970	1.030	1.185	0.935	**1.385**
→ETTh2	1.275	1.195	1.040	1.020	1.015	1.135	1.025	**1.370**
ESS (ours)	1.130	1.095	0.895	0.995	1.085	1.060	0.970	**1.275**
REDD	1.140	1.120	0.885	1.060	1.065	1.075	1.010	**1.340**
REFIT	1.160	1.170	0.880	1.040	0.990	1.085	0.965	**1.335**
Average	1.188	1.160	0.980	1.005	1.020	1.110	0.990	**1.340**

Use case of Archive

Forecasting Task. For this particular task, we benchmark generated data against various baselines, including DeepAR (Salinas et al. 2020), an autoregressive recurrent neural network; Diffusion denoising model D3VAE (Li et al. 2022), proposed recently used in Electricity AP dataset show good result against DeepAR; Transformers (Xu et al. 2020), improved time series prediction, we implement a Probabilistic Transformer (Tang and Matteson 2021) for this task. Our TIMEXAI toolkit provides a diverse selection of cutting-edge models suitable for energy forecasting purposes.

Dissagregation Task. To assess the effectiveness of the generated data in training Energy Disaggregation models, we utilize the sequence-to-sequence (S2S) (Chen et al. 2018b) and sequence-to-point (S2P) (Zhang et al. 2021a) as a baseline. This model combines CNNs and LSTM networks to capture temporal dependencies and is trained on both generated and real data. The performances of the model are evaluated on real data, providing insights into the quality and suitability of the generated data.

Training: Initially, we train selected models (T-Forcing, TimeVAE, GTGAN, TimeGAN, RCGAN, C-RNN-GAN, PSA-GAN, WaveGAN) on each dataset. GTGAN, TimeGAN, and RCGAN maintain similarity to real structure, while PSA-GAN and WaveGAN layers adapt for stability based on experiments. We then assess discriminators (comprising 5 GRU layers) and TCN-based encoders (Franceschi, Dieuleveut, and Jaggi 2019) for Discriminative and Context-FID scores. Our dataset archive might not optimally match tasks as indicated by Table **??** ratings, so benchmarking in the main paper gives priority to Dataport, Electricity AP, and the new TIMEXAI-ESS dataset for forecasting; REDD, REFIT, Uk-DALE, TIMEXAI-ESS for disaggregation. Evaluating

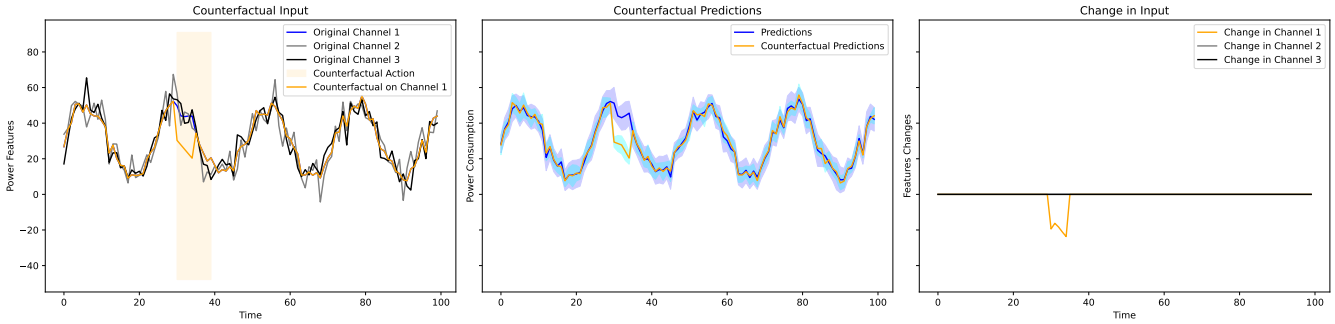

Figure 2: Visualizing t-SNE plots for different datasets, organized in rows. Left to right, the plots depict outcomes using different models. Each column provides the visualization for each of the 8 benchmarks. Magenta denotes actual data, and cyan denotes synthetic. PCA visualizations are provided in Appendix G of the Supplementary Materials.

optimal downstream model effectiveness involves training with Real+Generated data. Other dataset results are provided in the Appendix G of Supplement Materials.

Experimental Platform. The experiments are performed on four NVIDIA A40 GPUs and 40 Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz. The models are implemented in PyTorch.

Discussion

In this section, we discuss our findings on forecasting and disaggregation. Synthetic samples closely resemble original data, with GTGAN achieving the smallest deviation in Context-FID at 4.2%, while TimeVAE exhibits the highest deviation at 52.3%. PSA-GAN competes well with GTGAN, offering an initial exploration of GANs in forecasting. GT-GAN emerges as a robust sample generator for both tasks, consistently excelling in discrimination and prediction across datasets. C-RNN-GAN shows promise in forecasting for Uk-DALE and Dataport. Context-FID confirms these findings, and Figure 3 further demonstrates this alignment by revealing that GTGAN and PSA-GAN closely overlap with the real set. Models trained using a combination of real data and GTGAN outperform models trained solely on either real data or GTGAN individually (compare magenta and cyan rows in Table2). However, GTGAN and PSA-GAN show resilience with minor noise on TIMEXAI-ESS, due to ample, cleaned training data. TimeGAN encounters challenges with certain datasets. These results underscore GTGAN's superiority, particularly for Transformer ($\approx$ 32% better than DeepAR) and D3VAE ($\approx$ 37% better than DeepAR), while TimeGAN maintains competitiveness.

In disaggregation, Table 2 shows that GTGAN, PSA-GAN, and C-RNN-GAN compete, with TimeVAE lagging. TimeGAN's competitiveness with pristine data like TIMEXAI-ESS underscores its efficacy, slightly trailing GT-GAN and PSA-GAN. We also consider Figure 4, where GTGAN excels in short-term patterns but lags in capturing long-term trends. Notably, Figure 5 highlights TimeGAN's effectiveness. Our ablation study aims to illuminate these dynamics further.

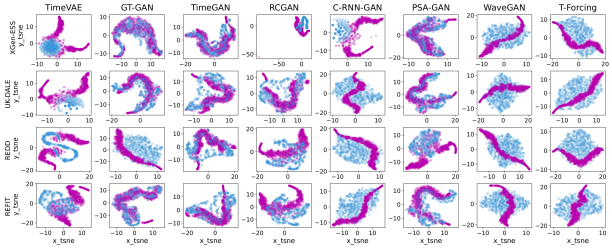

Figure 3: Visualizing t-SNE plots for different datasets, organized in rows. Left to right, the plots depict outcomes using different models. Each column provides the visualization for each of the 8 benchmarks. Magenta denotes actual data, and cyan denotes synthetic. PCA visualizations are provided in Appendix G of the Supplementary Materials.

Conclusions and future research

This paper introduces TIMEXAI, a pioneering open-source framework and archive that integrates both traditional and advanced time series generative models. TIMEXAI addresses the actionability and feasibility requirements for counterfactual explanations in time series prediction by proposing TIMEXAI, the first self-interpretable model that meets these criteria. Theoretical analysis demonstrates that TIMEXAI effectively identifies causal effects between time series inputs and outputs, even in the presence of confounding variables, thereby providing causal counterfactual explanations. Empirical results confirm that TIMEXAI achieves competitive prediction accuracy and generates actionable and feasible explanations. TIMEXAI also offers a comprehensive platform for reliable model training, data generation, and experiment tracking. Additionally, we introduce the TIMEXAI−ESS dataset to enhance disaggregation and forecasting tasks. A case study highlights GTGAN's superior performance on benchmark tasks, such as augmentation and missing value handling in energy time series. We anticipate that TIMEXAI and TIMEXAI−ESS will attract significant interest and benefit from continued community contributions.

Table 2: Predictive Score (MAE, RMSE) for disaggregation downstream tasks "Fridge consumption" (Lower is Better) ↓ indicates that lower values are better. Real data (magenta row) and fusion of real data and most valid generated case (cyan row). Values corresponding to best performance are bolded.

Metric	Method ↓	Uk-DALE		REFIT		REDD		TimeXAI-ESS	
	Baseline →	**S2S**	**S2S**	**S2S**	**S2S**	**S2S**	**S2S**	**S2S**	**S2S**
MAE ↓ (Lower the Better)	TimeVAE	.896±.088	.890±.001	.864±.001	.860±.002	.929±.007	.913±.010	.839±.030	.830±.010
	GTGAN	.254±.016	.251±.007	**.291±.002**	**.290±.004**	**.316±.012**	**.310±.003**	**.220±.024**	**.223±.025**
	TimeGAN	.628±.032	.621±.033	.622±.038	.617±.039	.640±.002	.634±.003	.538±.012	.537±.030
	C-RNN-GAN	.256±.040	.254±.012	.639±.012	.638±.013	.830±.030	.828±.010	.732±.102	.730±.003
	PSA-GAN	**.253±.001**	**.252±.002**	.330±.049	.331±.010	.331±.000	.328±.023	.330±.012	.325±.041
	WaveGAN	.286±.006	.282±.007	.643±.000	.642±.031	.336±.030	.332±.001	.237±.018	.231±.002
	T-Forcing	.757±.020	.754±.008	.741±.010	.740±.011	.833±.003	.829±.004	.624±.012	.619±.007
	Real	.198±.002	.194±.001	.193±.008	.188±.002	.186±.001	.183±.002	.161±.012	.156±.013
	Real+GTGAN	.197±.001	.186±.003	.193±.018	.193±.019	.196±.030	.196±.031	.163±.012	.163±.013
RMSE ↓ (Lower the Better)	TimeVAE	1.344±.012	1.344±.013	1.296±.008	1.291±.002	1.387±.012	1.387±.013	1.154±.142	1.154±.143
	GTGAN	.380±.018	.380±.019	**.316±.001**	**.314±.007**	**.375±.016**	**.373±.017**	**.380±.002**	**.378±.001**
	TimeGAN	.944±.060	.940±.001	.934±.002	.931±.003	.960±.036	.957±.002	.962±.005	.958±.001
	C-RNN-GAN	.484±.001	.477±.002	.452±.001	.451±.001	.485±.046	.486±.047	.352±.042	.382±.043
	PSA-GAN	**.381±.005**	**.379±.002**	.360±.002	.460±.013	.442±.004	.440±.005	.444±.012	.440±.003
	WaveGAN	.628±.000	.627±.001	.625±.001	.623±.002	.504±.000	.504±.001	.405±.018	.400±.001
	T-Forcing	.434±.027	.434±.028	1.001±.004	1.001±.005	1.251±.002	1.244±.003	1.250±.007	1.243±.003
	Real	.337±.008	.337±.001	.348±.015	.348±.016	.374±.013	.374±.014	.224±.006	.224±.007
	Real+GTGAN	.337±.004	.331±.009	.346±.003	.348±.001	.372±.001	.371±.004	.224±.001	.220±.002

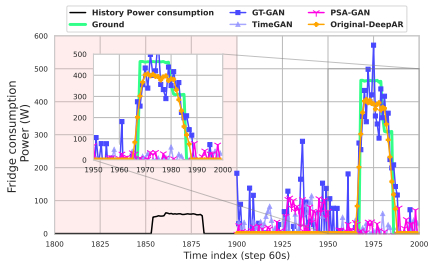

Figure 4: Forecasting performance on TimeXAI-ESS Test set (real) using DeepAR trained on generated and real TimeXAI-ESS Train set.

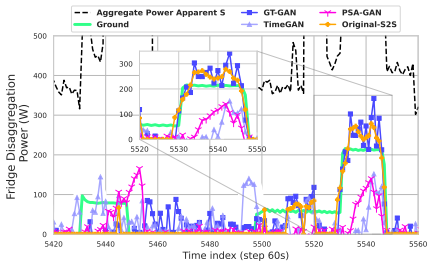

Figure 5: Dissagregation results on TimeXAI-ESS Test set (real) using S2S trained on generated and real TimeXAI-ESS Train set.

Reproducticibility Statement

The necessary information to replicate our experiments can be found in Supplementary Materials. Training specifics are outlined in Appendices A, B, C, and D, with additional experiment settings in Section F. Access the code we used is available at: https://TimeXAI.github.io/.

References

Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018a. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, 883–892. PMLR.

Chen, K.; Wang, Q.; He, Z.; Chen, K.; Hu, J.; and He, J. 2018b. Convolutional sequence to sequence non-intrusive load monitoring. *the Journal of Engineering*, 2018(17): 1860–1864.

Crabbé, J.; and Van Der Schaar, M. 2021. Explaining time series predictions with dynamic masks. In *ICML*, 2166–2177.

Demir, S.; Mincev, K.; Kok, K.; and Paterakis, N. G. 2021. Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Applied Energy*, 304: 117695.

Dhaou, A.; Bertoncello, A.; Gourvénec, S.; Garnier, J.; and Le Pennec, E. 2021. Causal and Interpretable Rules for Time Series Analysis. In *Proceedings of the 27th ACM SIGKDD*

Conference on Knowledge Discovery & Data Mining, 2764–2772.

Ding, H.; Ma, Y.; Deoras, A.; Wang, Y.; and Wang, H. 2022. Zero-shot recommender systems. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*.

Donahue, C.; McAuley, J.; and Puckette, M. 2018. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.

Enguehard, J. 2023. Learning Perturbations to Explain Time Series Predictions. In *ICML*, 9329–9342.

Esteban, C.; Hyland, S. L.; and Rätsch, G. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems*, 32.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.

Gupta, S.; Wang, H.; Lipton, Z.; and Wang, Y. 2021. Correcting exposure bias for link recommendation. In *International Conference on Machine Learning*, 3953–3963. PMLR.

Heo, J.; Lee, H. B.; Kim, S.; Lee, J.; Kim, K. J.; Yang, E.; and Hwang, S. J. 2018. Uncertainty-Aware Attention for Reliable Interpretation and Prediction. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Huang, H.; Wang, H.; and Mak, B. 2019. Recurrent poisson process unit for speech recognition. In *AAAI*, volume 33, 6538–6545.

Jeha, P.; Bohlke-Schneider, M.; Mercado, P.; Nirwan, R. S.; Kapoor, S.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2021. PSA-GAN: Progressive self attention GANs for synthetic time series.

Jeon, J.; KIM, J.; Song, H.; Cho, S.; and Park, N. 2022. GT-GAN: General Purpose Time Series Synthesis with Generative Adversarial Networks. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Letzgus, S.; Wagner, P.; Lederer, J.; Samek, W.; Müller, K.-R.; and Montavon, G. 2021. Toward Explainable AI for Regression Models. *arXiv preprint arXiv:2112.11407*.

Li, Y.; Lu, X.; Wang, Y.; and Dou, D. 2022. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022.

Lim, B.; Arik, S. O.; Loeff, N.; and Pfister, T. 2019. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv preprint arXiv:1912.09363*.

Lim, B.; Arık, S. Ö.; Loeff, N.; and Pfister, T. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764.

Lin, W.; Lan, H.; Wang, H.; and Li, B. 2022. OrphicX: A Causality-Inspired Latent Variable Model for Interpreting Graph Neural Networks. In *CVPR*.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Lundberg, S. M.; and Lee, S.-I. 2017a. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lundberg, S. M.; and Lee, S.-I. 2017b. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mao, C.; Chiquier, M.; Wang, H.; Yang, J.; and Vondrick, C. 2021a. Adversarial Attacks are Reversible with Natural Supervision. In *ICCV*.

Mao, C.; Gupta, A.; Cha, A.; Wang, H.; Yang, J.; and Vondrick, C. 2021b. Generative Interventions for Causal Learning. In *CVPR*.

Mogren, O. 2016. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*.

Natesan Ramamurthy, K.; Vinzamuri, B.; Zhang, Y.; and Dhurandhar, A. 2020. Model agnostic multilevel explanations. *Advances in neural information processing systems*, 33: 5968–5979.

Nemirovsky, D.; Thiebaut, N.; Xu, Y.; and Gupta, A. 2022. CounteRGAN: Generating counterfactuals for real-time recourse and interpretability using residual GANs. In Cussens, J.; and Zhang, K., eds., *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, 1488–1497. PMLR.

Pan, Q.; Hu, W.; and Chen, N. 2021. Two Birds with One Stone: Series Saliency for Accurate and Interpretable Multivariate Time Series Forecasting. In *IJCAI*, 2884–2891.

Pawlowski, N.; Coelho de Castro, D.; and Glocker, B. 2020. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33: 857–869.

Plumb, G.; Molitor, D.; and Talwalkar, A. S. 2018. Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31.

Queen, O.; Hartvigsen, T.; Koker, T.; He, H.; Tsiligkaridis, T.; and Zitnik, M. 2023. Encoding Time-Series Explanations through Self-Supervised Model Behavior Consistency. In *NeurIPS*.

Rajapaksha, D.; and Bergmeir, C. 2022. LIMREF: Local Interpretable Model Agnostic Rule-based Explanations for Forecasting, with an Application to Electricity Smart Meter Data. *arXiv preprint arXiv:2202.07766*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016a. " 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, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016b. " 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*, 1135–1144.

Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shi, J.; Stebliankin, V.; and Narasimhan, G. 2023. The Power of Explainability in Forecast-Informed Deep Learning Models for Flood Mitigation. *arXiv preprint arXiv:2310.19166*.

Shi, Y.; Srey, P.; and Tsang, I. ???? Dynamic-Aware GANs: Time-Series Generation with Handy Self-Supervision.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.

Siddiqui, S. A.; Mercier, D.; Munir, M.; Dengel, A.; and Ahmed, S. 2019. Tsviz: Demystification of deep learning models for time-series analysis. *IEEE Access*, 7: 67027–67040.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sirojan, T.; Phung, B. T.; and Ambikairajah, E. 2018. Deep Neural Network Based Energy Disaggregation. In *2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE)*, 73–77.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Tang, B.; and Matteson, D. S. 2021. Probabilistic Transformer For Time Series Analysis. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Tonekaboni, S.; Joshi, S.; Campbell, K.; Duvenaud, D. K.; and Goldenberg, A. 2020. What went wrong and when? Instance-wise feature importance for time-series black-box models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 799–809. Curran Associates, Inc.

Vahdat, A.; and Kautz, J. 2020. NVAE: A Deep Hierarchical Variational Autoencoder. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wang, H. 2017. *Bayesian Deep Learning for Integrated Intelligence: Bridging the Gap between Perception and Inference*. Ph.D. thesis, Hong Kong University of Science and Technology.

Wang, H.; Mao, C.; He, H.; Zhao, M.; Jaakkola, T. S.; and Katabi, D. 2019. Bidirectional inference networks: A class of deep bayesian networks for health profiling. In *AAAI*, volume 33, 766–773.

Wang, H.; Wang, N.; and Yeung, D. 2015. Collaborative deep learning for recommender systems. In *KDD*, 1235–1244.

Wang, H.; and Yeung, D.-Y. 2016. Towards Bayesian deep learning: A framework and some existing methods. *TDKE*, 28(12): 3395–3408.

Wang, H.; and Yeung, D.-Y. 2020. A Survey on Bayesian Deep Learning. *CSUR*, 53(5): 1–37.

Wang, S.; Zhou, T.; and Bilmes, J. 2019. Bias also matters: Bias attribution for deep neural network explanation. In *International Conference on Machine Learning*, 6659–6667. PMLR.

Wang, Y.; Menkovski, V.; Wang, H.; Du, X.; and Pechenizkiy, M. 2020. Causal discovery from incomplete data: A deep learning approach.

Wang, Z.; Yan, W.; and Oates, T. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, 1578–1585. IEEE.

Xu, H.; Liu, Q.; Xiong, D.; and van Genabith, J. 2020. Transformer with depth-wise lstm. *arXiv preprint arXiv:2007.06257*.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6199–6203. IEEE.

Yoon, J.; Jarrett, D.; and Van der Schaar, M. 2019. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32.

Zhang, J.; Sun, J.; Gan, J.; Liu, Q.; and Liu, X. 2021a. Improving Domestic NILM Using An Attention-Enabled Seq2Point Learning Approach. In *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 434–439. IEEE.

Zhang, Y.; Tiňo, P.; Leonardis, A.; and Tang, K. 2021b. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5): 726–742.